Inquire: Large-scale Early Insight Discovery for Qualitative Research

Pablo Paredes[1,3], **Ana Rufino Ferreira**[1], **Cory Scillaci**[1], **Gene Yoo**[1], **Pierre Karashchuk**[1],
Dennis Xing, Coye Cheshire[2], **John Canny**[2]
[1]Electrical Engineering and Computer Science Department, University of California, Berkeley, USA
[2]Information School, University of California, Berkeley, USA
[3]Computer Science Department, Stanford University
paredes@cs.stanford.edu, {peparedes, ana, jfc}@eecs.berkeley.edu
{schillaci, geneyoo, pierre, yxing, coye, canny}@berkeley.edu

ABSTRACT

We introduce *Inquire*, a tool designed to enable qualitative exploration of utterances in social media and large-scale texts. As opposed to keyword search, *Inquire* allows the effective use of sentences as queries to quickly explore millions of documents to retrieve semantically-similar sentences. We apply *Inquire* to LiveJournal.com (LJ) database, which contains millions of personal diaries, and we use semantic embeddings trained in LJ or Google News (GN) datasets. We present the system design through iterative evaluations with qualitative researchers. We show how queries become a part of the inductive process, enabling researchers to try multiple ideas while gaining intuition and discovering less-obvious insights. We discuss the choice of LJ as a rich source of public posts, the preference for GN embeddings which link formal language (e.g. "*reminiscence triggers*") with colloquial expressions (e.g. "*music brings back memories*"), the interplay between tool and user, and potential qualitative and social research opportunities.

Author Keywords

Qualitative research, semantic, keyword search, big data, text data, large corpus, insights, exploratory, hypothesis formation.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

Despite the growing popularity of qualitative research in CSCW and HCI, tools designed to draw early insights from large online corpuses remain mostly unexplored. Often, the burden of accessing "excessive" amounts of text data for qualitative analysis is handled by selective sampling of texts, or simply reducing the scope of the research. When a corpus is small, such as a few dozen interviews or texts, exploration is not especially difficult to manage. However, for a large corpus—the equivalent of thousands or millions of interviews, field notes, or diaries—this task becomes difficult if not impossible without computational support.

A significant number of researchers in the CSCW and HCI communities work with textual social media data. From online dating platforms [32] to online diaries [30], qualitative researchers in CSCW and HCI regularly access and analyze enormous quantities of texts. While dozens or even hundreds of interviews may have at once seemed large, the rise of social media and online information sharing tools now raises the potential for millions of units of qualitative analysis.

One common set of tools used in qualitative research are coding, analysis and search packages such as HyperResearch [37], Nvivo [34], MaxQDA [35], and AtlasTi [36]. These tools are often used by loading texts into a database and manually reading through them, or using word-matching tools to find relevant content. The main drawback of such systems is the time required to generate relevant results. Since these tools were primarily designed for coding and manual analysis, insight mining of large texts could be prohibitively time consuming or perhaps even impossible.

Today, the sheer quantity of text data made available online often leads to a compromise between more traditional qualitative text analysis, and automated coding procedures that are often quantitative in nature. Recently, systems leveraging computational social science [4,10] and social computing [13,14] have shown novel ways to draw meaning from large-scale text corpuses. Most of these tools however, offer little ability for the researcher to observe and influence the computational process, or even hide the raw data by providing only summarizing terms or statistics.

In this paper, we introduce *Inquire*, a novel instrument designed to enable the early insight discovery process across millions of independent units of text. One of the goals of the *Inquire* tool is to balance the aforementioned need for manual qualitative insight generation with the ability to find potentially relevant passages in a veritable sea of text. To accomplish this, we focus on allowing the researcher to "see" the data in its original form and leverage modern Natural Language Processing (NLP) algorithms that facilitate the retrieval of data. With the *Insight* tool, we aim to maintain the central, reflective role of the researcher in the theme-generation process, while helping them more easily explore, find, identify, and expose semantically related passages.

Semantic relevancy is defined based on algorithms that capture word correspondence such as "*paris*" and "*france*", "*london*" and "*england*", etc. [23] Semantic embeddings that maximize accuracy are 'deep learned' by training over samples of a word and its surrounding text in a very large text corpus such as Google News (GN). Algorithms such as word2vec [21] trained on GN have shown success in generating such embeddings, allowing compositionality operations such as "*king*" - "*man*" + "*woman*" = "*queen*" [22]. We take advantage of these properties to perform queries at the sentence level.

Rather than developing automated solutions, or generating abstractions of the data, we propose an exploratory process that enables the interplay between large corpuses of data and the user via the formulation of simple, yet fast and scalable queries (Figure 1). A query can be as simple as a generic term or word, or as complex as a sentence, or a group of sentences. In fact, answers are often used as inputs, contributing to the refinement of the research idea. Our initial prototype uses GN embeddings to perform queries on Livejournal.com (LJ). The choice of LJ and GN as the underlying datasets was carefully analyzed. On the one hand, LJ represents a deep and dense dataset that carries ideoscincracies, it is anonymus, eclectic, unstructured, and therefore a good approximation of the blogosphere. On the other hand, GN carries information that links more formal (journalistic) writing with colloquial personal expressions. This combination proved to be relevant to the formulation of meaningful queries among researchers.

We contrast our approach with Internet Search Engines (SE) such as Google or Bing, commonly used during early exploration of a research topic. Keyword queries can be used to retrieve info from many texts, including public Internet sites such as LJ. This allows researchers to access content that the underlying algorithms, such as "PageRank", determine as relevant, based on assumptions about link popularity and term frequency scores [26]. Of critical importance, however, is the fact that less-obvious insights are not necessarily found in the top hits returned by these tools. This is not due to any flaw in typical SE tools, but rather a different priority: search algorithms prioritize finding typical and popular documents; in contrast, qualitative researchers are often looking for insightful content that may or may not be typical at all.

In summary, we focus our efforts on building an instrument, rather than a fully automated system, analog to building a "piano" rather than a "radio". Sometimes researchers are interested in consuming what was prepared to you, (e.g. music in the radio), but this is at the expense of exploring their creativity and find what works for their own purposes (e.g. playing the piano). *Inquire* in this case represents the "piano", leaving it to researchers to interpret and decide what is most relevant. We argue that there are at least four advantages of our system over more traditional, manual exploration, search, and theme-discovery approaches:

Economy. Our approach is very economical when compared to manual text analysis methods. Researchers can explore different topics within a large corpus in minutes through the use of a text-based interface with simple controls and filters. With little practice, this allows the researcher to retrieve relevant info and to try different query techniques.

Intuition Development. The ability to rapidly explore or exploit several ideas aids to the researcher's ability to iteratively modify queries. Such constant refinement, adds to the reflective process, which in turn supports the development of good intuition that leads to insight discovery.

Neutral Representations of Similarity and Diversity. Manual exploration and coding of qualitative text data often requires the researcher to consider one's own biases and assumptions when analyzing texts from different respondents, sources, etc. In our approach, relevant themes are extracted based on semantic methods that are impartial to the source.

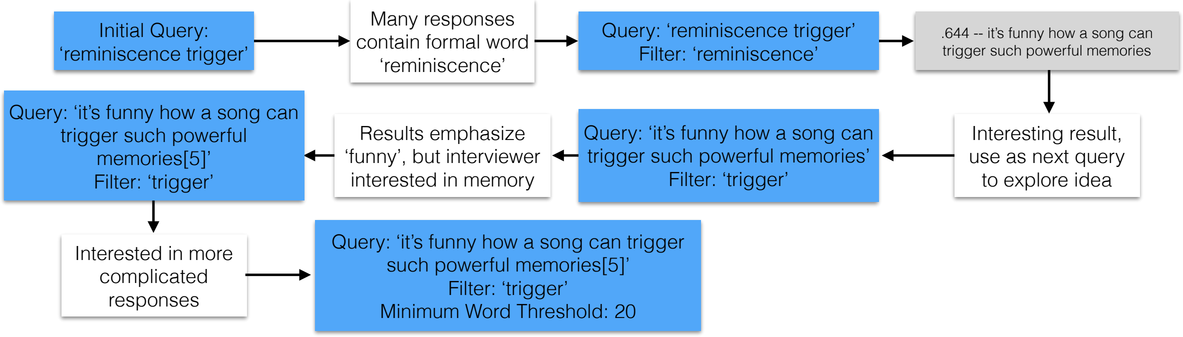

Initial Query: 'reminiscence trigger'
Many responses contain formal word 'reminiscence'
Query: 'reminiscence trigger' Filter: 'reminiscence'
.644 -- it's funny how a song can trigger such powerful memories
Interesting result, use as next query to explore idea
Query: 'it's funny how a song can trigger such powerful memories' Filter: 'trigger'
Results emphasize 'funny', but interviewer interested in memory
Query: 'it's funny how a song can trigger such powerful memories[5]' Filter: 'trigger'
Interested in more complicated responses
Query: 'it's funny how a song can trigger such powerful memories[5]' Filter: 'trigger' Minimum Word Threshold: 20

Figure 1: Example, based on a real query, how the queries evolve in response to the returned answers.

Scalability. In some cases, the effectiveness of our method is commensurate with the size of the corpus and the quality of the embeddings. With the current trend towards rapid progress in information retrieval scalability and accuracy of semantic embedding methods, this means that our approach will improve in usefulness along with these developments.

The paper offers a thorough description of the system's rationale and design process and how the researcher's feedback informed the feature selection. We close with a discussion of some insights that could inform the way similar tools could evolve in the future. Finally, we provide a description of some samples of future research in the fields of psychology, privacy and wellbeing.

BACKGROUND AND RELATED WORK

Fundamentals of Qualitative Coding Methods

Qualitative research methods often rely on a critical epistemological assumption that the researcher must get as close as feasible to the object of study (often other humans) to learn about and understand subjective evidence based on individual views and experiences [5]. In stark contrast to most quantitative methods (e.g, experiments, surveys), qualitative researchers knowingly bring their own values to a study by admitting and actively stating these values and biases as a way to position oneself in the research [5].

Qualitative research heavily depends on the *authenticity* of people's experiences more than the size of the sample [28]. Thus, the purpose of coding and analyzing rich, qualitative interviews, historical records, and observations is not to create probabilistic generalizations of a large sample to a defined population. In contrast, qualitative researchers use semi-structured interviews and other forms of qualitative data to draw out and analyze personal experiences in order to understand social phenomena in a nascent research area [28]. In doing so, qualitative researchers are able to highlight concepts, experiences, attitudes, beliefs and behaviors that (1) illustrate behaviors and attitudes that inform our understanding of the topic, and (2) help us build a stronger connection between relevant theories and experiences.

An essential part of the qualitative research process is *coding* the data. In qualitative inquiry, a code is, *"a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data"* [27]. Coding is the link between data collection and explanations of meaning [3]. Coding is typically a manual, labor-intensive process where one or more researchers carefully reads, watches, or listens to all of the data in order to tag and select specific passages, instances, or utterances for similarity of meaning. The qualitative analysis process includes coding, categorizing, and ultimately creating concepts and themes from the data [27]. While codes deal with the short phrases that describe salient content within the data, themes and concepts represent the higher-level, abstract ideas that aid in the development of theory. There are numerous methods for coding, categorizing and creating themes in qualitative data—and all of them involve manual, iterative processes where the researcher is the research 'instrument' [27,28].

Qualitative Data Analysis Tools

In recent years, Qualitative Data Analysis (QDA) tools have been quite helpful to streamline qualitative research. Tools such as hyperRESEARCH [37], MaxQDA [35], Atlas.ti [36], and N-Vivo [34] are among the most popular. These tools provide a wealth of functions for researchers to code small to medium-size datasets quickly and efficiently. These packages help generate insights drawn from different data such as videos, text, annotations and recordings. While most qualitative software allows for keyword and synonym search, these packages lack semantic analysis modules. They also do not have direct knowledge connections to sources such as social media or libraries. Thus, one of the outcomes for our system is the ability to work in concert with existing qualitative software, perhaps as a complementary plug-in.

Keyword Search

Keyword searches were the first approach to mining a large corpus of data for relevant information. Exploration techniques using single-word queries date back to keyword-in-context indexing (KWIC) [11,17] which provides snippets of text surrounding the search query and a key for reading the text in its original context. This model is still relevant, but the results are difficult to parse and recent work has focused on visualization methods such as Word Tree [31] which aggregates results with identical word sequences and displays them in a tree-like structure. Later visual text exploration tools based on the Word Tree visualization paradigm include WordSeer [24] which makes it easier to navigate a corpus by facilitating switching between different views, adding summary statistics and applying filters (such as date ranges or other metadata). Other refinements included imposing specific grammatical requirements on the results [25]. Modern algorithms such as PageRank from Google take into consideration the importance of a document, based on the number of links to other important documents and term frequency metrics [26].

A standard goal of semi-structured interviews is to obtain non-obvious insights, responses pertinent to the original topic but not expected by the interviewer. Most such interview responses are not expected to explicitly contain words from the interviewer's questions. For this purpose, keyword search techniques can be too restrictive. Some efforts have been made to search a corpus directly by document level themes. One such example provides a flexible visualization and exploration tool [9] based on topic modeling, a technique to identify latent themes across a collection of documents [1]. As an alternative to the latent concept approach, the relatedness of texts has also been computed using explicit semantic analysis on natural concepts as defined by humans via Wikipedia [33]. An- other attempt to leverage the Wikipedia corpus for semantic annotation of content is described in [12,20].

Semantic Similarity

Researchers have made efforts to developing a semantic similarity metric applicable at the word or sentence level. Schemes for measuring similarity at the word level include LSI and Point-wise Mutual Information and Information Retrieval (PMI-IR) [29], both of which learn semantic relation- ships based on co-occurrence of words in a training corpus. More recently, distributed word embeddings learned with recurrent neural networks have become state of the art through algorithms like word2vec. Word2vec semantic similarity refers to recognizing word correspondences, such as "paris" and "france", "London" and "England", etc. In the original work by Mikolov, et. al [23], they explore different language model architectures, such as Continuous Bag of Words (CBOW) and continuous skip-gram. The former predicts the current word based on context (surrounding text), while the latter predicts surrounding words given the current word. We use the skip-gram variation of word2vec, which enables us to search for semantically similar expressions to our queries. For example, while searching for "*illness*" you can find expressions such as *"My chest hurts"*. In the system design section we describe how we combine word embeddings to search at the sentence level (Figure 1).

Building on these efforts, various models to compute short text similarity based on word-level semantic similarity were also developed [15,19]. A recent paper also attempts to develop direct embeddings at the sentence level using long short term memory (LSTM) neural networks, called skip-thought vectors [16]. We argue that the use of these technologies provides a much richer form of matching than traditional keyword-based search and even first-generation semantic search techniques such as Latent Semantic Indexing (LSI) [8]. We bolster this argument with targeted evaluations where we show that perceived quality of the results was improved by filtering out surface matching (i.e. keyword-matching) posts from semantically matching posts.

Social Media Monitoring

In the past decade, many tools to monitor and analyze social media have emerged. Companies such as Sysomos [38], Hootsuite [39], and Mblast [40] provide a variety of services to help track brand perception online. They tend to focus on monitoring many social media sites in real-time, keyword search over these sites, geographical and temporal trends for various keywords, ranking influence of people, and sometimes sentiment analysis. Few, if any, social media analytics companies offer a semantic search similar to our proposed design. As such, we believe that our tool may provide a complementary look at the social media data, beyond what is currently offered today.

In the next two sections we describe the way how drawbacks and advantages of the preceding technologies affected the design rationale and system choices.

DESIGN MOTIVATION AND RATIONALE

Our system originates from social researchers' need to study complex topics composed of multiple schemas, like understanding life events, mindsets, emotions, nutritional behaviors, etc. Such topics are rarely well defined, fairly unstructured and require qualitative exploration. Many researchers are drawn to explore large corpora of existing data in search for information and insights to frame their research. Concretely, our early motivation came from the need to analyze stressful life-events that are associated with poor health outcomes. Our first challenge was to choose a database that would have a good balance between positive and negative life events. We picked Livejournal.com (LJ), a social media site where users maintain an anonymous personal, publicly accessible online journal or diary, with themes including diverse areas such as *health*, *lifestyle*, *hobbies*, and many others. Soon, we realized that parsing through LJ with traditional tools such as regular expressions or even keyword search algorithms such as PageRank was not enough to extract relevant information. Building queries that are broad enough using these constrained methods proved too complex and time consuming. This is when we decided to focus our research in adopting modern NLP techniques that enable semantic searching capabilities.

We selected word2vec [21,23] as the main building block for our system. As described in the background section, this is a neural semantic embedding method that performs well at a variety of semantic tasks. The true power of the technique is that the resulting feature space demonstrates *semantic* structure. For example, vec("*king*") − vec("*man*") + vec("*woman*") has a greater cosine similarity to vec("*queen*") than to the vector of any other word in the dictionary. We can also generalize this comparison technique beyond the scope of single-words. By performing vector addition on the word vector embeddings of corresponding words and normalizing the resulting sum, we can compare the semantic proximity of complete sentences. This procedure is described in the System Components section.

We build our tool based on word2vec embeddings from Google News (GN) and apply it to LJ posts, which tend to be significantly longer than other social media sites, and therefore provide a good initial example of the challenges associated with analyzing large numbers of large texts. Within the LJ corpus there are significant tracts of text where users discuss attitudes, values and personal experiences around various topics. We argue that these tracts are similar to the diversity and breadth of responses that researchers would analyze in other large collections of text-based qualitative data, including combinations of semi-structured interviews, field notes, personal histories, and other information-rich texts on the Internet. Furthermore, LJ is a denser and deeper than commonly used datasets such as Twitter; in many ways it can be seen as an antithesis of Twitter. We estimated that social science researchers will be able to find relevant and rich data in LJ.

Early in the design process we realized the power of GN embeddings. We contrasted embeddings trained on the LC corpora and embeddings trained on GN corpora. The latter presented the best way to link between the colloquial nature of LJ users and a more formal language used by researchers. GN embeddings work as a bridge between the way researchers think about complex concepts and the way people write about them in LJ. Although we leverage the peculiarities of LJ and GN, *Inquire's* workflow was envisioned as a tool to be used with other corpora and other semantic embeds such as LSTM. Finally, we wanted to enable very fast searching, with the purpose of creating a fluid process of inquire without the need to self-censor or filter any questions due to time burdens.

In the next sections we provide detailed information about the system structure to hopefully motivate other researchers to replicate our system to be used with other data sets.

SYSTEM COMPONENTS

We process the LJ dataset into matrices that are later used to perform fast computations allowing the generation of a cosine distance representation of semantic similarity. We introduce each of the elements of our system's workflow (Figure 2) in the following subsections.

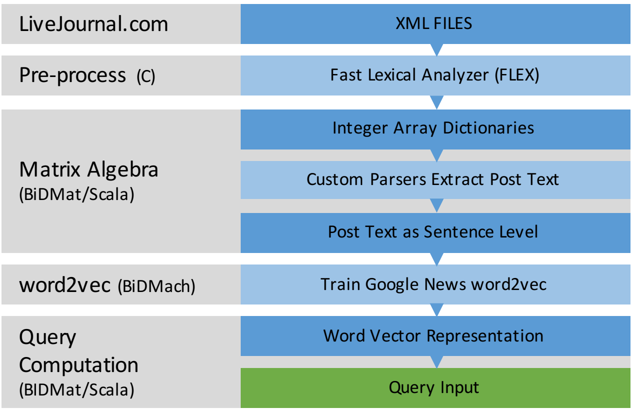
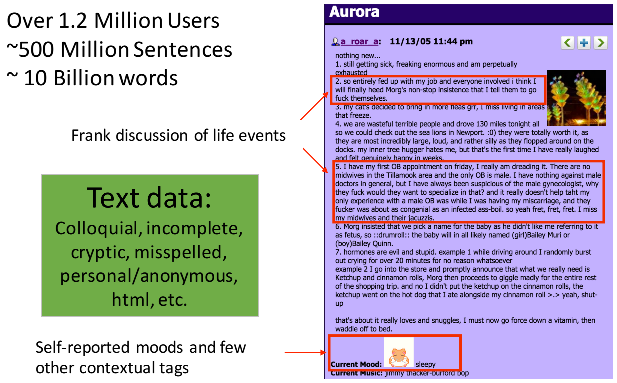

LiveJournal.com
XML FILES
Pre-process (C)
Fast Lexical Analyzer (FLEX)
Matrix Algebra
(BiDMat/Scala)
Integer Array Dictionaries
Custom Parsers Extract Post Text
Post Text as Sentence Level
word2vec (BiDMach)
Train Google News word2vec
Query
Computation
(BIDMat/Scala)
Word Vector Representation
Query Input

Figure 2: Inquire workflow

LiveJournal Corpus

For this study, we obtained a data set consisting of all public English-language LJ posts (Figure 3) as of November 2012 in raw XML format. As previously mentioned, LiveJournal is a social networking service where users can keep a blog, journal, or diary. Users have their own journal pages, which show all of their most recent journal entries. Each journal entry can also be viewed on its own web page which includes comments left by other users. As of 2012, LJ in the US received about 170 million page views each month from 10 million unique visitors. Our data contains about 1 million users with a total of 64,326,865 text posts. The median sentence length is 10 words and, as expected, the distribution of sentence lengths follows a power law. Of the users that provided their date of birth, the majority were in the 17-25 age group. Users who chose to identify their location were primarily in the US (72%), with significant populations in Russia, Canada, the UK, and Australia. Additionally, users were able to indicate their binary gender; of those who did so 45% identified as male and 55% as female.


Over 1.2 Million Users
~500 Million Sentences
~ 10 Billion words
Frank discussion of life events
Text data:
Colloquial, incomplete, cryptic, misspelled, personal/anonymous, html, etc.
Self-reported moods and few other contextual tags
Aurora

a_roar_a: 11/13/05 11:44 pm

nothing new...
1. still getting sick, freaking enormous and am perpetually exhausted
2. so entirely fed up with my job and everyone involved i think I will finally heed Morg's non-stop insistence that I tell them to go fuck themselves.
3. my cat's decided to bring in more fleas grr, I miss living in areas that freeze.
4. we are wasteful terrible people and drove 130 miles tonight all so we could check out the sea lions in Newport. :0) they were totally worth it, as they are most incredibly large, loud, and rather silly as they flopped around on the docks. my inner tree hugger hates me, but that's the first time I have really laughed and felt genuinely happy in weeks.
5. I have my first OB appointment on friday, I really am dreading it. There are no midwives in the Tillamook area and the only OB is male. I have nothing against male doctors in general, but I have always been suspicious of the male gynecologist, why they fuck would they want to specialize in that? and it really doesn't help taht my only experience with a male OB was while I was having my miscarriage, and they fucker was about as congenial as an infected ass-boil. so yeah fret, fret, fret. I miss my midwives and their jacuzzis.
6. Morg insisted that we pick a name for the baby as he didn't like me referring to it as fetus, so ::drumroll:: the baby will in all likely named (girl)Bailey Muri or (boy)Bailey Quinn.
7. hormones are evil and stupid. example 1 while driving around I randomly burst out crying for over 20 minutes for no reason whatsoever
example 2 I go into the store and promptly announce that what we really need is Ketchup and cinnamon rolls, Morg then proceeds to giggle madly for the entire rest of the shopping trip. and no I didn't put the ketchup on the cinnamon rolls, the ketchup went on the hot dog that I ate alongside my cinnamon roll >.> yeah, shut-up

that's about it really loves and snuggles, I must now go force down a vitamin, then waddle off to bed.

Current Mood: sleepy
Current Music: jimmy thacker-burford bop

Figure 3: LiveJournal Dataset (~10B words / ~500M sentences)

Data preprocessing

We started with a raw XML dataset which consists of more than 9 billion words in over 500 million sentences. In order to apply (NLP) algorithms, we implemented a data pipeline that allows us to extract and clean the post text in a computationally efficient form.

The first step was to tokenize the corpus using FLEX (Fast Lexical Analyzer), a scanner which maps each word in the corpus to a unique number and saves the mapping in a dictionary. For efficiency, we keep only the 994,949 most common words and discard the rest (all of which occur fewer than 41 times in the entire dataset). Care is taken to properly tokenize numbers and emoticons such as ":-)", "=<" or ">:P".

After removing non-textual content such as images, hyperlinks, and XML formatting data, a bag-of-words (BOW) representation is saved for each sentence from the posts as a column in a sparse feature matrix. The rows of this matrix correspond to the entries in the dictionary above, and the values in the matrix are the counts of each entry for each sentence. For each sentence, we also store the ID of the post it belongs to and the user who wrote it so that the original post can be found online. Finally, we keep the tokenized sentence data to display results when performing the query.

Matrix Algebra Platform

Inquire was implemented and executed in the Scala language and more concretely in the BID DATA suite of libraries [2]: BiDMach –an optimized library for machine learning and BidMat – an optimized library for matrix computations. The system run in Linux. We used a single node with 64GB of RAM and four HD units. With this system were able to run up to 1.2% of the data in less than 1 second. Using threaded memory, we were able to run 100% of the data in 5 (five) minutes. We estimate that in order to run 100% of the data in real time (i.e. <10 seconds) we could use about 10 machines with similar characteristics to the single node.

Word2vec Embeddings

We obtained semantic information using word2vec [22,23], a word embedding model used in a variety of NLP applications, including analogy tasks, sentence completion, machine translation [18], and topic modeling [7,10]. Word embedding models map each dictionary word into a lower dimensional continuous vector representation. With word2vec, this embedding is learned automatically from a sample corpus using a recurrent neural network.

We utilize two embedding models trained on two different corpuses. First, we trained our own embedding model on the corpus of LJ posts using the skip-gram with negative sampling (SGNS) implementation (as opposed to CBOW) of word2vec as recommended in [22]. This model gives 300-dimensional embeddings for the 994,949 words in our dictionary. Conveniently, Mikolov et. al published a pre-trained SGNS word2vec model for open source access [21]. This model was trained on a 100 billion-word GN corpus, and gives 300-dimensional embeddings for 3 million unique words. The terms from our LJ dictionary that are not included in this model are mapped to zero vectors. Common stop words (i.e. "the", "an", "who") provide little semantic information and are also mapped to zero vectors.

Query Design

Our design goal is to perform semantic searches on the dataset from individual words or full sentences. To do this, we find the similarity of sentences based on the embeddings of their individual words. Specifically, we define a sentence embedding to be the normalized mean of its constituent word vector embeddings. The query is also converted to a sentence embedding and tested on the data set using cosine similarity. Note that multi-sentence queries are also possible by treating the entire query in the same way. Word embeddings have been employed in more sophisticated approaches, which outperform the cosine similarity method when used in short text strings [15]. However, our approach has sufficient power to favor rapid retrieval from a large corpus.

Semantic querying uses word2vec embeddings trained on GN or LJ datasets. Each embedding is trained with the same master dictionary, to guarantee that each index corresponds to the same word. Let **w2vMat ∈ R300 × #words** be the word2vec embedding, where **#words** is the number of words in the master dictionary, and **300** is the chosen vector dimension. The featurized sentence data is the bag of words matrix **dataMat ∈R#words × #sents**, where **#sents** is the number of sentences considered from the LiveJournal dataset. The sentence data is converted into vectors:

$$\mathbf{dataVec} = \mathbf{w2vMat} * \mathbf{dataMat}$$

where matrix **dataVec** is further normalized along each column, and represents the numerical matrix where we perform the querying. Querying is performed by converting the original query into a bag of words (**queryWords ∈ R#words × 1**), and then calculating its corresponding vector

$$\mathbf{queryVec} = \mathbf{w2vMat} * \mathbf{queryWords}$$

which is then normalized. The semantic relationship is determined by cosine distance, which is the dot product of **queryVec** with the columns magic, since they are both normalized. The following array, **distMat ∈ R1 × #sents**, represents the semantic score between the query vector and the sentence vectors from LiveJournal data:

$$\mathbf{distMat} = \mathbf{queryVec} * \mathbf{dataVec}$$

While the system above gives us valuable info, some additional filters proved very helpful to refine interesting responses (Table 1).

Feature	Use	Analogy
Minimum Sentence Length	Pick only sentences with a min number of words	Short versus expanded versions.
Filter sentences containing words	Show sentences with semantic equivalence	Use different ways to explain a topic.
Individual word weights	Increase the importance of a word within a sentence	Focus on one element of an answer
Original link	Show original link where the sentence appeared.	Provide context for a specific statement.

Table 1: Special Features for query manipulation

Based on our interaction with the researchers we set the minimum number of words to 7. However, the user can adjust this parameter at any time. We also allow the user to filter out responses containing a specific word (or a set of words). Finally, we give users the ability to add an *"importance*" weighting to words in the search query. In this case, when calculating the embeddings of the query sentence, we perform a weighted mean of the individual word embeddings. Note that these weights can be positive or negative. In the latter case, responses with semantic relation to the negatively weighted word are suppressed. To understand the context of the sentences returned by the query, we also provide the user a link to the original post on the LJ site (Figure 4).

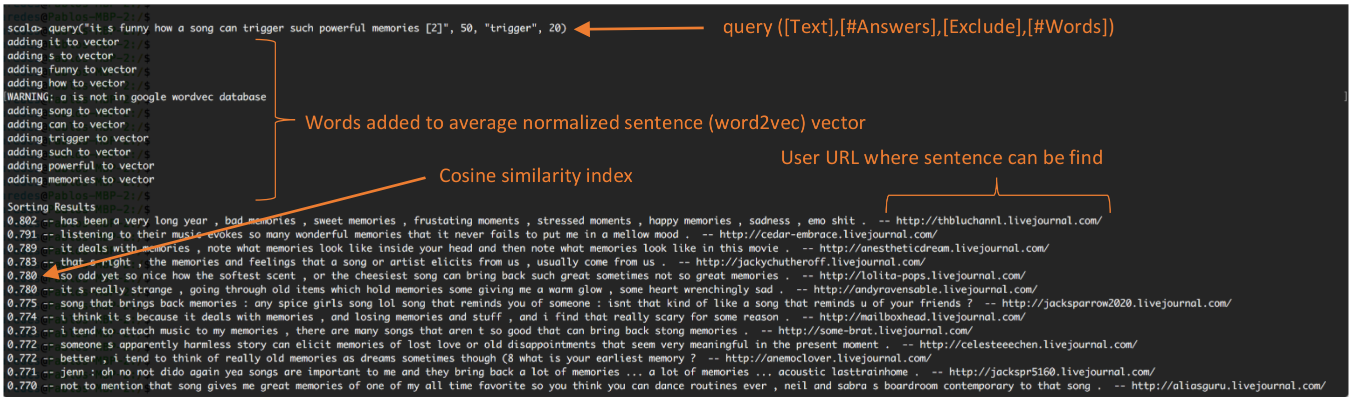
Text-based Interface

Our user interface is a text-based command tool using Scala, and running on the BID Data platform [2]. To load the interface we predefine the size of data and the source of the embeddings (LJ or GN). Once loaded, we can run queries in the command line (Figure 4). To do this we use the function:

query([Text],[#Answers],[Exclude],[#Words])

Text:	query phrase
#Answers:	number of answers shown
Exclude:	remove answers containing this word
#Words:	show sentences with min number of words

We took advantage of the shell history to retrieve earlier queries and modify them. Researchers often copied and pasted answers as input for new queries. Opening a link was performed by right clicking the URL.


query ([Text],[#Answers],[Exclude],[#Words])
Words added to average normalized sentence (word2vec) vector
User URL where sentence can be find
Cosine similarity index

```
scala> query("it s funny how a song can trigger such powerful memories [2]", 50, "trigger", 20)
adding it to vector
adding s to vector
adding funny to vector
adding how to vector
WARNING: a is not in google wordvec database
adding song to vector
adding can to vector
adding trigger to vector
adding such to vector
adding powerful to vector
adding memories to vector

Sorting Results
0.802 -- has been a very long year , bad memories , sweet memories , frustating moments , stressed moments , happy memories , sadness , emo shit .  -- http://thbluchannl.livejournal.com/
0.791 -- listening to their music evokes so many wonderful memories that it never fails to put me in a mellow mood .  -- http://cedar-embrace.livejournal.com/
0.789 -- it deals with memories , note what memories look like inside your head and then note what memories look like in this movie .  -- http://anestheticdream.livejournal.com/
0.783 -- that s right , the memories and feelings that a song or artist elicits from us , usually come from us .  -- http://jackychutheroff.livejournal.com/
0.780 -- so odd yet so nice how the softest scent , or the cheesiest song can bring back such great sometimes not so great memories .  -- http://lolita-pops.livejournal.com/
0.780 -- it s really strange , going through old items which hold memories some giving me a warm glow , some heart wrenchingly sad .  -- http://andyravensable.livejournal.com/
0.775 -- song that brings back memories : any spice girls song lol song that reminds you of someone : isnt that kind of like a song that reminds u of your friends ?  -- http://jacksparrow2020.livejournal.com/
0.774 -- i think it s because it deals with memories , and losing memories and stuff , and i find that really scary for some reason .  -- http://mailboxhead.livejournal.com/
0.773 -- i tend to attach music to my memories , there are many songs that aren t so good that can bring back stong memories .  -- http://some-brat.livejournal.com/
0.772 -- someone s apparently harmless story can elicit memories of lost love or old disappointments that seem very meaningful in the present moment .  -- http://celesteeechen.livejournal.com/
0.772 -- better , i tend to think of really old memories as dreams sometimes though (8 what is your earliest memory ?  -- http://anemoclover.livejournal.com/
0.771 -- jenn : oh no not dido again yea songs are important to me and they bring back a lot of memories ... a lot of memories ... acoustic lasttrainhome .  -- http://jackspr5160.livejournal.com/
0.770 -- not to mention that song gives me great memories of one of my all time favorite so you think you can dance routines ever , neil and sabra s boardroom contemporary to that song .  -- http://aliasguru.livejournal.com/
```

Figure 4: Inquire Command Line Interface

ITERATIVE SYSTEM DEVELOPMENT

Methodology

We focused on academic researchers with experience on coding and exploring text-based data. We interviewed nine researchers, five of them were PhDs and the rest were PhD students. Four of them were women and five man. They represented a variety of fields with different research projects at various research stages (Table 2). Most of them focused on a sample population that can generalize universally to humankind. Only a couple of them required particular populations for their research. We described LJ as library of anonymous "*diaries*" covering mostly mundane topics. Researchers were free to choose any research topic and spent 45 minutes interacting with the tool. In order to reduce learning curve effects and to concentrate in the searching flow, the intervener executed the queries upon their command. They were introduced to different features after a few queries, on average every 5 minutes.

Early Idea Validation

We approached researchers early in the system design process to gain perspective on qualitative methods and tools. We interviewed a social worker [R1] and two ethnographers [R2, R3]. We asked their opinion on the exploration of a large-scale corpus. The social worker [R1] believed that such a tool would be beneficial to frame a "*universal*" perspective of people's thoughts. She believed that a fundamental need was to have "*context*," i.e. to be able to modify the unit of analysis from sentences to posts or journals and back.

> *If you can search in all those diaries would be really good… and if there are many that are similar it helps to contrast.*

Both ethnographers [R2, R3] wanted a tool allowing them to discover a group of people with common interests would be valuable. [R2] provided an important insight. He mentioned that "*saturation*" of responses (repeatedly receiving similar responses) could be informative in qualitative research. So, we decided against implementing any kind of diversity filter.

> *For us seeing that something is repeated many times is valid. There is a concept in ethnography called "saturation", and we use it all the time.*

First iteration: Desktop Use (0.1% of corpus)

We started with a prototype using 0.1% of the data (~5M sentences representing around 12K users) running in a desktop computer with 16GB of RAM. Running queries in just a few seconds allowed researchers to make meaningful connections between queries and answers. They found themselves running multiple queries with incremental variations many times. We began interviewing a public health researcher [R4] focused on how people maintain a low-weight after dieting. He quickly found value in some of the sentences. However, he wanted access to the surrounding sentences or preferably the entire post.

> *If you can show me some context would be good, like, maybe the sentence before, or even better before and after. Yeah, the link to the page would be good also, …*

During the process the researcher asked if it was possible to weight the words. He noted that sometimes ancillary words dominated the answers. We later implemented this feature in the final iteration (Table 1).

ID	Research Domain	Topic	Population Scope	Research Stage	Design Phase
R1	Social Welfare	Gender Studies	Particular	Writing Results	Early Idea Validation
R2	Information Science	Wikipedia	Universal	Writing Results	Early Idea Validation
R3	Ethnographer	Data Science Research	Universal	Conceptualization	Early Idea Validation
R4	Public Health	Nutrition	Universal	Case Finding	Lo-Fi / Mid-Fi Prototype
R5	Information Science	Airbnb	Universal	Validation	Lo-Fi
R6	Sociologist	Family Relationships / Emotions	Universal	Validation	Lo-Fi / Mid-Fi Prototype
R7	Anthropologist	Migration Patterns	Particular	Validation	Refused to Participate
R8	HCI	Social Activism / Reminiscence	Universal	Case Finding / Validation	Mid-Fi Prototype
R9	Lawyer / Sociologist	Law Accessibility	Universal	Conceptualization	Mid-Fi Prototype

Table 2: Set of researchers who tried system. Some participated in more than one evaluation iteration.

We found no sentences including the word “Airbnb”, because our dataset cut date was November 2012. Despite this limitation she explored queries concerning issues between neighbors. We noted how the researcher moved from keyword search towards using answers as queries.

> *… now, this is better, but you have to make some “crazy” assumptions to get to this level.*

As closing remarks, she mentioned that it was important for the user to become a “*smart user*”. The process of creating queries helped her form new ideas around the research topic. On the other hand, the perils and frustration of generating *"good"* queries should be addressed by having relevant data, if possible with a description of the broad topics available.

> *For the smart system vs smart user, I was just thinking along the lines of how well the system is equipped to predict what the user means (think of early Altavista vs Google today) and how that affects how easy it is for the user to get what they want…*

We finalized our evaluations with a sociologist focused on online social relationships [R6]. He was well versed in search models, and he wanted to have an initial description of the system. He started with a query about “*family holidays*” (Figure 5). He appreciated seeing many expressions similar to the query as well as others semantically close but syntactically different. He visited the user pages (Figure 6), going back and forth between the sentences and full posts. He noticed that some sentences came from the same user. He felt that the tool was helping him to find *“genres”* of users.

> *This is impressive...the actual responses for a query are very similar … Ah!, if you can search for all those “authors”, people that write about the same topic you could get something like “genres” … yeah, this would be very useful*

He was impressed to find phrases that were relevant but which did not include the keywords. While searching for “*family holidays*” he found results with and without the word "*family*". He believed that a valuable feature would be to filter query words from the results. We implemented this in the final iteration (Table 1).

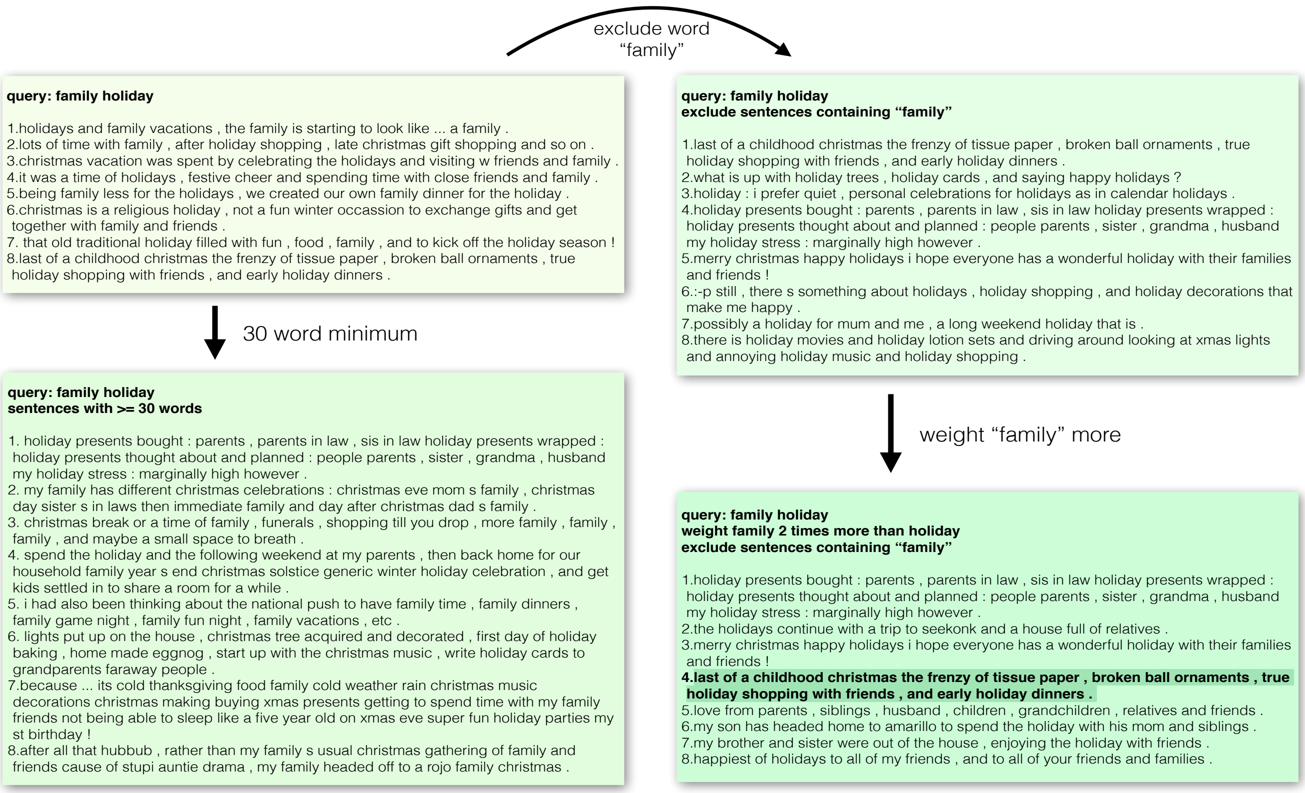
> *… more than searching for the tail, it would be better to filter out “words” like “family”, which comes many times… and see only things like [this one] about the “brother and the sister”?*

Second-Iteration: Specialized Server (1.2% of corpus)

We used a 64GB server to search over 1.2% of the data (~60M sentences representing ~150K users). The public health researcher [R4] tried now his query: *“I lost weight and kept it off”* and found some excellent case studies.

> *...and I think I will send this to my colleagues in the CDC, we had been looking for cases like these ones for quite some time*

He suggested that we filter answers by words - similar to [R6]. He also wanted to see the surrounding sentences by default. Finally, he said that he would like to select journals he intends to work on and perform additional search on those journals alone.



query: family holiday

1.holidays and family vacations , the family is starting to look like ... a family .
2.lots of time with family , after holiday shopping , late christmas gift shopping and so on .
3.christmas vacation was spent by celebrating the holidays and visiting w friends and family .
4.it was a time of holidays , festive cheer and spending time with close friends and family .
5.being family less for the holidays , we created our own family dinner for the holiday .
6.christmas is a religious holiday , not a fun winter occassion to exchange gifts and get together with family and friends .
7. that old traditional holiday filled with fun , food , family , and to kick off the holiday season !
8.last of a childhood christmas the frenzy of tissue paper , broken ball ornaments , true holiday shopping with friends , and early holiday dinners .

query: family holiday
sentences with >= 30 words

1. holiday presents bought : parents , parents in law , sis in law holiday presents wrapped : holiday presents thought about and planned : people parents , sister , grandma , husband my holiday stress : marginally high however .
2. my family has different christmas celebrations : christmas eve mom s family , christmas day sister s in laws then immediate family and day after christmas dad s family .
3. christmas break or a time of family , funerals , shopping till you drop , more family , family , family , and maybe a small space to breath .
4. spend the holiday and the following weekend at my parents , then back home for our household family year s end christmas solstice generic winter holiday celebration , and get kids settled in to share a room for a while .
5. i had also been thinking about the national push to have family time , family dinners , family game night , family fun night , family vacations , etc .
6. lights put up on the house , christmas tree acquired and decorated , first day of holiday baking , home made eggnog , start up with the christmas music , write holiday cards to grandparents faraway people .
7.because ... its cold thanksgiving food family cold weather rain christmas music decorations christmas making buying xmas presents getting to spend time with my family friends not being able to sleep like a five year old on xmas eve super fun holiday parties my st birthday !
8.after all that hubbub , rather than my family s usual christmas gathering of family and friends cause of stupi auntie drama , my family headed off to a rojo family christmas .

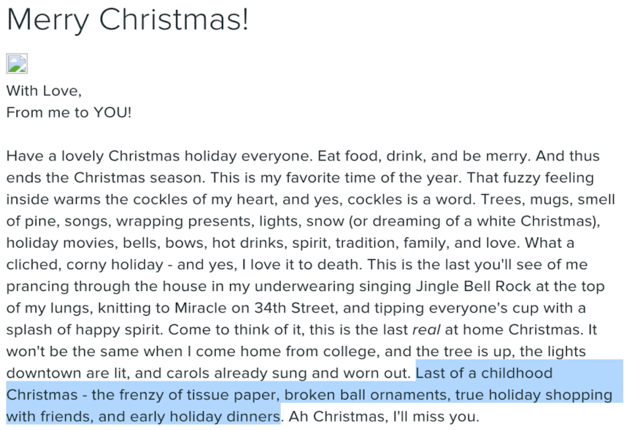
query: family holiday
exclude sentences containing “family”

1.last of a childhood christmas the frenzy of tissue paper , broken ball ornaments , true holiday shopping with friends , and early holiday dinners .
2.what is up with holiday trees , holiday cards , and saying happy holidays ?
3.holiday : i prefer quiet , personal celebrations for holidays as in calendar holidays .
4.holiday presents bought : parents , parents in law , sis in law holiday presents wrapped : holiday presents thought about and planned : people parents , sister , grandma , husband my holiday stress : marginally high however .
5.merry christmas happy holidays i hope everyone has a wonderful holiday with their families and friends !
6.:-p still , there s something about holidays , holiday shopping , and holiday decorations that make me happy .
7.possibly a holiday for mum and me , a long weekend holiday that is .
8.there is holiday movies and holiday lotion sets and driving around looking at xmas lights and annoying holiday music and holiday shopping .

query: family holiday
weight family 2 times more than holiday
exclude sentences containing “family”

1.holiday presents bought : parents , parents in law , sis in law holiday presents wrapped : holiday presents thought about and planned : people parents , sister , grandma , husband my holiday stress : marginally high however .
2.the holidays continue with a trip to seekonk and a house full of relatives .
3.merry christmas happy holidays i hope everyone has a wonderful holiday with their families and friends !
4.last of a childhood christmas the frenzy of tissue paper , broken ball ornaments , true holiday shopping with friends , and early holiday dinners .
5.love from parents , siblings , husband , children , grandchildren , relatives and friends .
6.my son has headed home to amarillo to spend the holiday with his mom and siblings .
7.my brother and sister were out of the house , enjoying the holiday with friends .
8.happiest of holidays to all of my friends , and to all of your friends and families .

Figure 5: Example for que query: “*Family Holiday*” and the effect of filters in the query results on the full data.



Merry Christmas!

With Love,
From me to YOU!

Have a lovely Christmas holiday everyone. Eat food, drink, and be merry. And thus ends the Christmas season. This is my favorite time of the year. That fuzzy feeling inside warms the cockles of my heart, and yes, cockles is a word. Trees, mugs, smell of pine, songs, wrapping presents, lights, snow (or dreaming of a white Christmas), holiday movies, bells, bows, hot drinks, spirit, tradition, family, and love. What a cliched, corny holiday - and yes, I love it to death. This is the last you'll see of me prancing through the house in my underwearing singing Jingle Bell Rock at the top of my lungs, knitting to Miracle on 34th Street, and tipping everyone's cup with a splash of happy spirit. Come to think of it, this is the last *real* at home Christmas. It won't be the same when I come home from college, and the tree is up, the lights downtown are lit, and carols already sung and worn out. Last of a childhood Christmas - the frenzy of tissue paper, broken ball ornaments, true holiday shopping with friends, and early holiday dinners. Ah Christmas, I'll miss you.

Figure 6: Public post on LiveJournal.com, corresponding to the highlighted result in Figure 2.

Finally, an anthropologist [R7] was not interested the tool due to its lack of demographics info. However, she believed she could use the tool at very early stages of her research. She asked if we could mine demographic data from posts.

Final Iteration: Filter Implementation

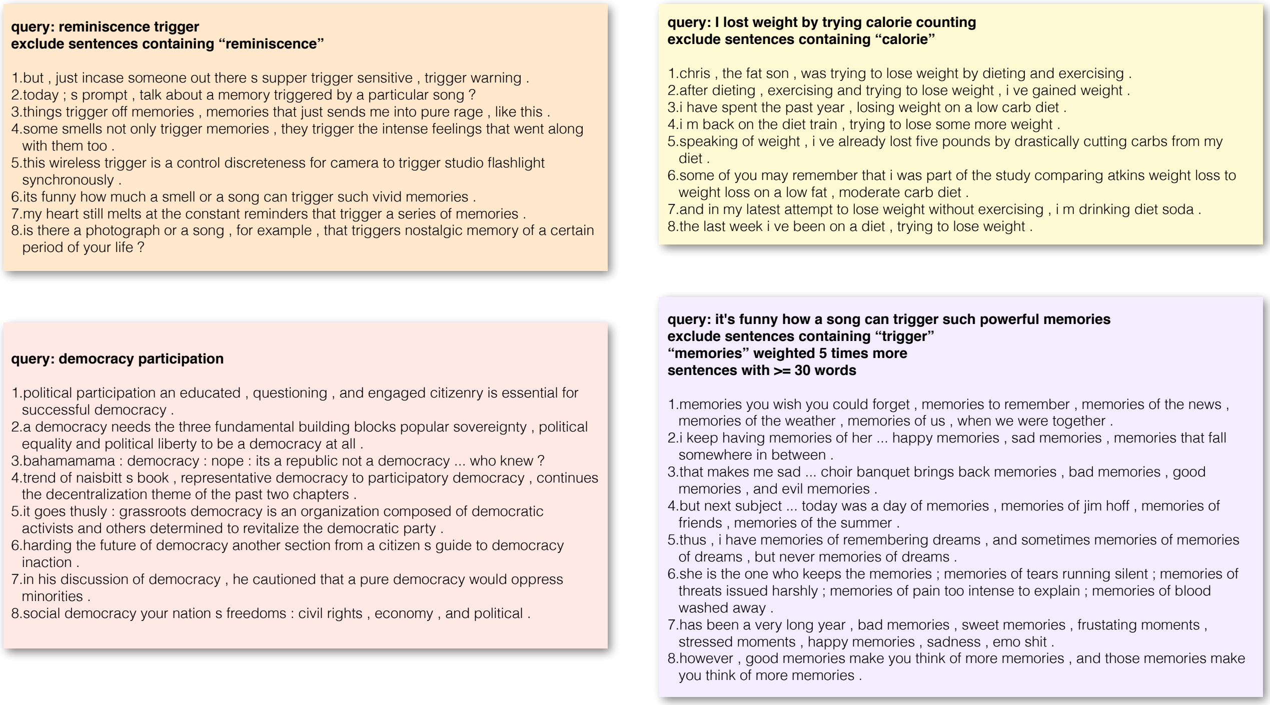
The sociologist [R6] tested his queries with new filters (Table1).

> *It is a really good tool to be able to find the chunks of data that I want to use in my studies and I would feel pretty comfortable in being able to use that [word2vec]*

An HCI researcher [R8] studying *online activism* and *multimedia tools for elder people* found little merit in the data for *online activism*, perhaps for being a highly specialized topic. However, he found interesting references once he adapted his search strategy from keywords to colloquial expressions such as "*please sign my petition*". He found useful to move away from traditional keyword queries towards more *"meaningful"* expressions.

> *I think the problem might be when you think about a search or a query tool we are way too much biased by our daily search experience, "all about keywords" but it seems to me that this, for this to work, you have to give an example.*

He was impressed with the results for "*reminiscence trigger*". He quickly found expressions around *songs* and *smells,* which took him a long time to discover and validate. He believed that this tool has the potential to support research at a formative phase as well as during refinement of the hypothesis. Later we showed off-line results from 100% of the data (Figure 7). He found these results compelling, but did not improved the value he gained by using 1% of the data. We closed our evaluations with a lawyer and sociologist [R9] researching *access to justice*. The initial results were peculiar because of the phrasing of her queries. When searching for "*presence in court*" many answers were about *tennis courts*. However, when we added the filter to eliminate the word "*ball*" she found many interesting quotes about the judiciary system. She believed that the tool could save money and time instead of going "*door to door*" at the early stages.



query: reminiscence trigger
exclude sentences containing "reminiscence"

1.but , just incase someone out there s supper trigger sensitive , trigger warning .
2.today ; s prompt , talk about a memory triggered by a particular song ?
3.things trigger off memories , memories that just sends me into pure rage , like this .
4.some smells not only trigger memories , they trigger the intense feelings that went along with them too .
5.this wireless trigger is a control discreteness for camera to trigger studio flashlight synchronously .
6.its funny how much a smell or a song can trigger such vivid memories .
7.my heart still melts at the constant reminders that trigger a series of memories .
8.is there a photograph or a song , for example , that triggers nostalgic memory of a certain period of your life ?

query: I lost weight by trying calorie counting
exclude sentences containing "calorie"

1.chris , the fat son , was trying to lose weight by dieting and exercising .
2.after dieting , exercising and trying to lose weight , i ve gained weight .
3.i have spent the past year , losing weight on a low carb diet .
4.i m back on the diet train , trying to lose some more weight .
5.speaking of weight , i ve already lost five pounds by drastically cutting carbs from my diet .
6.some of you may remember that i was part of the study comparing atkins weight loss to weight loss on a low fat , moderate carb diet .
7.and in my latest attempt to lose weight without exercising , i m drinking diet soda .
8.the last week i ve been on a diet , trying to lose weight .

query: democracy participation

1.political participation an educated , questioning , and engaged citizenry is essential for successful democracy .
2.a democracy needs the three fundamental building blocks popular sovereignty , political equality and political liberty to be a democracy at all .
3.bahamamama : democracy : nope : its a republic not a democracy ... who knew ?
4.trend of naisbitt s book , representative democracy to participatory democracy , continues the decentralization theme of the past two chapters .
5.it goes thusly : grassroots democracy is an organization composed of democratic activists and others determined to revitalize the democratic party .
6.harding the future of democracy another section from a citizen s guide to democracy inaction .
7.in his discussion of democracy , he cautioned that a pure democracy would oppress minorities .
8.social democracy your nation s freedoms : civil rights , economy , and political .

query: it's funny how a song can trigger such powerful memories
exclude sentences containing "trigger"
"memories" weighted 5 times more
sentences with >= 30 words

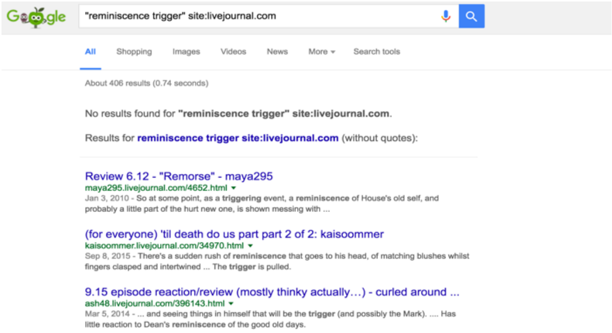
1.memories you wish you could forget , memories to remember , memories of the news , memories of the weather , memories of us , when we were together .
2.i keep having memories of her ... happy memories , sad memories , memories that fall somewhere in between .
3.that makes me sad ... choir banquet brings back memories , bad memories , good memories , and evil memories .
4.but next subject ... today was a day of memories , memories of jim hoff , memories of friends , memories of the summer .
5.thus , i have memories of remembering dreams , and sometimes memories of memories of dreams , but never memories of dreams .
6.she is the one who keeps the memories ; memories of tears running silent ; memories of threats issued harshly ; memories of pain too intense to explain ; memories of blood washed away .
7.has been a very long year , bad memories , sweet memories , frustating moments , stressed moments , happy memories , sadness , emo shit .
8.however , good memories make you think of more memories , and those memories make you think of more memories .

Figure 7: Example queries and results ran on 100% of the data, and with distinct filter

> *I believe that I can use this tool in two moments. In the beginning of my research when I want to decide which issue I will talk about say I will try to figure out what is going on, what people are talking about, what are the main issues that are going on. I can bring one issue that is interesting to help me make the right question. It's nice to see what people are talking [about]. I can then build the question to be tested. And then I see [it] another time to use it when I want to make research with people and I don't have enough money or don't have enough time to go to the field, …*

System Benchmark

We contrasted our final iteration with a traditional keyword searching tool, Google search with the modifier "*source:livejournal.com*". Figure 8 shows the results rendered by a Google using the PageRank algorithm for the same query "reminiscence trigger" used in Figures 1 and 7.



About 406 results (0.74 seconds)

No results found for **"reminiscence trigger" site:livejournal.com**.

Results for **reminiscence trigger site:livejournal.com** (without quotes):

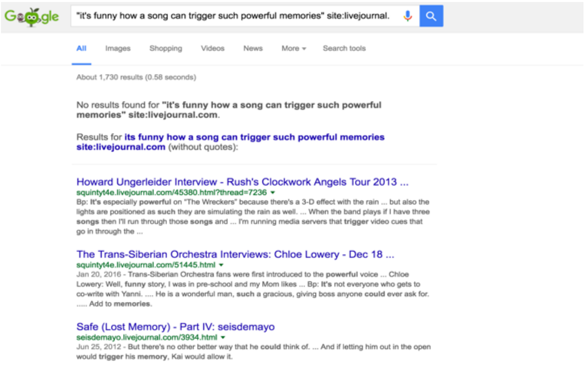
Review 6.12 - "Remorse" - maya295
maya295.livejournal.com/4652.html
Jan 3, 2010 - So at some point, as a **triggering** event, a **reminiscence** of House's old self, and probably a little part of the hurt new one, is shown messing with ...

(for everyone) 'til death do us part part 2 of 2: kaisoommer
kaisoommer.livejournal.com/34970.html
Sep 8, 2015 - There's a sudden rush of **reminiscence** that goes to his head, of matching blushes whilst fingers clasped and intertwined ... The **trigger** is pulled.

9.15 episode reaction/review (mostly thinky actually…) - curled around ...
ash48.livejournal.com/396143.html
Mar 5, 2014 - ... and seeing things in himself that will be the **trigger** (and possibly the Mark). .... Has little reaction to Dean's **reminiscence** of the good old days.

Figure 8: Google results for query "reminiscence trigger site:livejournal.com"

Google did not find results for the exact match "reminiscence trigger" but did find results containing both keywords. However, when accessing the links, words were in different sentences, or in various posts. Additionally, traversing the results was quite cumbersome. Google presents only 10 results per page, and moving through sequential pages is time consuming. We then tried a sentence query - commonly used in *Inquire* - *"It's funny how a song can trigger such powerful memories"* (Figure 9).


"it's funny how a song can trigger such powerful memories" site:livejournal.
All
Images
Shopping
Videos
News
More
Search tools

About 1,730 results (0.58 seconds)

No results found for **"it's funny how a song can trigger such powerful memories" site:livejournal.com**.

Results for **its funny how a song can trigger such powerful memories site:livejournal.com** (without quotes):

Howard Ungerleider Interview - Rush's Clockwork Angels Tour 2013 ...
squintyt4e.livejournal.com/45380.html?thread=7236
Bp: **It's** especially **powerful** on "The Wreckers" because there's a 3-D effect with the rain ... but also the lights are positioned as **such** they are simulating the rain as well. ... When the band plays if I have three **songs** then I'll run through those **songs** and ... I'm running media servers that **trigger** video cues that go in through the ...

The Trans-Siberian Orchestra Interviews: Chloe Lowery - Dec 18 ...
squintyt4e.livejournal.com/51445.html
Jan 20, 2016 - Trans-Siberian Orchestra fans were first introduced to the **powerful** voice ... Chloe Lowery: Well, **funny** story, I was in pre-school and my Mom likes ... Bp: **It's** not everyone who gets to co-write with Yanni. .... He is a wonderful man, **such** a gracious, giving boss anyone **could** ever ask for. ..... Add to **memories**.

Safe (Lost Memory) - Part IV: seisdemayo
seisdemayo.livejournal.com/3934.html
Jun 25, 2012 - But there's no other better way that he **could** think of. ... And if letting him out in the open would **trigger** his **memory**, Kai would allow it.

Figure 9: Google results for query "It's funny how a song can trigger such powerful memories site:livejournal.com"

Again, quotes did not help, but Google did find documents containing all the keywords. In contrast to the same query in *Inquire* (Figure 7), Google fails to capture the meaning of the query. It does identify documents that have the keywords, but they are random collections of information that add no value to the query at hand. We tried other queries represented in this paper and got very similar results across the board.

Discussion

As exposed in this paper, Inquire presents a novel perspective for early insight discovery with large-scale text. Mainly we observe a departure from using individual keywords into using phrases or expressions. This shift was celebrated by some researchers who found it a closer representation of their train of thought from a qualitative perspective. Some researchers, more inclined to use keywords, quickly understood the concept when they started reusing answers as queries.

Additionally, researchers were able to create new queues on the fly, or drill down existing ones by replacing words, applying filters or adding together different answers. This fast manipulation of the queries did not only generated better results, it improved the way researchers search. Researchers mentioned that while becoming a "smart searcher" they were also reflecting and framing their research ideas better. Bringing together qualitative researchers with powerful NLP algorithms into a single interactive tool presents itself as an excellent opportunity for research. We argue that HCI and CSCW researchers should invest into creating such instruments to leverage researchers' creativity and analysis abilities while using large corpuses of data.

Another valuable insight came from the use of embeddings that came from a dataset with different properties. By using GN's embeddings, we were able to link formal words such as "reminiscence" to colloquial expressions such as "it's funny how a song can trigger powerful memories". This helped the researcher open up to a vocabulary that reveals the way people think when writing freely. These observations led the researchers to explore queries that were a mix of keywords and colloquial expressions. On the other hand, LJ embeddings were observed to work better once the researcher tried to use only "colloquial" expressions. In general, the choice of GN or LJ generated different answers. We are in the process of developing new algorithms such as LSTM [16] to train on the LJ dataset. In LSTM syntactic order matters. We show a glimpse of this approach in the future work section.

Merging answers into a query is a way to perform a human-guided "clustering" algorithm, which reduces a large universe of data into a specific set of relevant answers. This clustering ability was welcomed by researchers who used less-obvious answers as queries. For example, by looking at similar answers, the researcher studying family relationships proposed the existence of "genres" of intra-family behaviors.

The ability to expand the level of abstraction from a sentence to a webpage gave access to contextual info which sometimes was as, or even more, interesting than the actual answer itself. The researcher studying nutrition found users that were exhaustively recording their calorie counts. He found such cases worth studying in more detail. We are currently exploring ways of expanding the way we use and present contextual data back to the user.

Finally, the ability to mine through data that was written at free-will reduces biases compared to more curated content. By observing free expressions of self-documentation, self-inquiry or introspection, researchers can discover ideas that may not appear in more traditional methods which suffer from different degrees of interviewer biases. This of course, can lead to proposing new theories or lines of research.

In the future work section, we describe some efforts already in place, where we are using the LJ dataset to discover mindsets (i.e. thoughts that represent a simplification of reality) about topics such as food, privacy and exercise.

Limitations

This system is limited to the data and meta data provided by the underlying corpus. In the case of LJ we have limited demographics data. It is possible to detect the country of the user, but no Socio-Economic data is available. Furthermore, our dataset has a cut of date that precludes the research of newer topics such Airbnb. This can be overcome by using fresh data, which we are in the process of obtaining. However, we argue that this system could still form general concepts to guide research focused in particular populations.

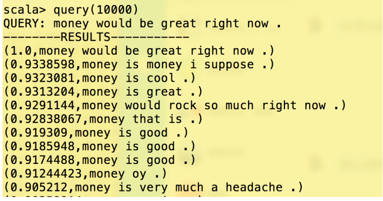
Another limitation is the lack of knowledge of the underlying topics present in the data. Clustering, topic modeling or other techniques would improve the possibility for researchers to determine beforehand if the underlying dataset would be conducive to valid results. This was evident on the researcher looking for *Airbnb* could have determined that the LJ dataset was not the right one. Undoubtedly, there is value in topic modeling of a dataset, however there is also a risk of creating a selection bias that would make the researcher either limit the use of the tool or chose specific datasets for different research topics. We argue that a dataset like LJ carries a broad mix of mundane topics with various degrees of depth.

FUTURE WORK

We are currently supporting social science projects while we develop a UI for the tool. We want to test novel embeddings, scrapping new data, performing real time search and visualizing data. Beyond exploration, we also want to experiment with generation of content. We describe each of these future system improvements below.

Additional Semantic Technologies.

Word2vec is a fast and moderately effective semantic embedding scheme. But it ignores word order and phrase structure. State-of-the-art methods use recurrent neural nets (RNNs) such as LSTM (Long Short-Term Memory) [16]. New algorithms model local and global sentence structure, and can be used to model paragraph and document-level structure. We tried an initial implementation of LSTM on the corpus using an existing sentence as query (Figure 10).



```
scala> query(10000)
QUERY: money would be great right now .
--------RESULTS-----------
(1.0,money would be great right now .)
(0.9338598,money is money i suppose .)
(0.9323081,money is cool .)
(0.9313204,money is great .)
(0.9291144,money would rock so much right now .)
(0.92838067,money that is .)
(0.919309,money is good .)
(0.9185948,money is good .)
(0.9174488,money is good .)
(0.91244423,money oy .)
(0.905212,money is very much a headache .)
```

Figure 10: LSTM-based results to sentence queries in LJ

As shown, there is a lot of similarity in the structure of the sentences. This may be useful in some analysis where syntactic content matters. For example, mining for "mindsets", which tend to have a concrete syntactic shape: "[N] is [A]". It remains a research challenge to determine what are the best types of embeddings to be used for different research challenges.

Larger-Scale Live Matching.

Semantic matching is expensive at the full scale of Livejournal. With 500 million sentences and 300-dimensional embedding, one needs 600 GB of "hot" data in memory to do model matching. That requires either custom hardware or a cluster of machines. Both techniques require modern parallel computing technique to be effective. Another option is to generate some efficient search method incorporating some knowledge about the embeddings. We can use Locality Sensitive Hashing (LSH) [6] to leverage a fast index to retrieve nearby sentences.

Topic Modeling and Labeling

We plan to apply different topic modeling algorithms to characterize the LJ dataset or others. Additionally, we plan to use novel topic modeling systems such as Empath [10] to label or categorize answers on the fly. We started testing this feature by labeling sentences related to food mindsets. We observed up to 74% accuracy in single dimension binary labeling (healthy versus unhealthy food). However, in two dimensions (health/unhealthy x indulgent/depriving) we observed only a high accuracy rate of about 93% for true negatives (sentences classified as "*other*") but a low 18% accuracy for true positives.

Synthetic Text Generation.

One powerful affordance of the LSTM design is the generation of text as well as query matching. That is, one can produce fully synthetic user output in response to a query. This synthetic text could be used for brainstorming, or as high quality input for translations. While one loses the ability to explore a user's post in context, one gains diversity and representativeness of the synthetic posts (via controls on the variety of their synthesis) and higher levels of privacy protection relative to retrieving true posts. The quality and utility of these posts is very much an open question.

UI and Visualization

We are exploring UI designs as well as ways to visualize jointly raw data, semantic similarity scores, and contextual information. One potential use will be to visualize topics dominating the data set and how the query results fall in different meta categories. We also want to visualize the position where data fall on the long-tail of data to inform the user what piece of data they are observing and analyzing. We acknowledge that adding a thorough UI and visualization could enhance the benefits and uses of *Inquire*

Applied Research

We expect to work more with social science and qualitative researchers that are using it to "mine" for complex constructs. One of them is mining for "mindsets", which are high level thought simplifications of reality. So far researchers have found value by observing the many ways people think about food, and are starting to formulate new theories around food mindsets. *Inquire* is therefore already enhancing the insight discovery phase in actual qualitative research projects.

CONCLUSION

In this paper, we introduced *Inquire*, a new tool for qualitative insight mining applied to a large-scale, colloquial text corpus. We take advantage of advances in semantic word embeddings to provide a versatile search tool based on researchers performing many fast queries phrased as sentences. Features for modifying and augmenting the search results were designed with the feedback of several qualitative researchers. Their responses highlight the tool's potential advantages in inductive thinking, and insight mining based on large datasets. *Inquire* enables powerful interaction between machine and user with the aim to make the researcher the ultimate driver of knowledge, instead of trying to automate and simplify data.

Inquire bridges the divide between formal research questions and a large collection of informal anecdotes. As stated by a researcher, this tool is unlike others in that it is "*not all about keywords*". Instead, the researcher states that he "*can also profit from Inquire by simply thinking of examples of expressions to search on*". Compared to traditional keyword searching, *Inquire* helps explore results whose connection to the original query are initially not obvious to the researcher and which ultimately lead to valuable insights about a topic.

ACKNOWLEDGMENTS

REFERENCES

1. David M. Blei. 2012. Introduction to Probabilistic Topic Modeling. *Communications of the ACM* 55: 77–84. http://doi.org/10.1145/2133806.2133826

2. John Canny and Huasha Zhao. 2013. Big Data Analytics with Small Footprint : Squaring the Cloud. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*: 95–103. http://doi.org/10.1145/2487575.2487677

3. K Charmaz. 2001. Grounded theory. *Contemporary Field Research*, 335–352.

4. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*: 3267. http://doi.org/10.1145/2470654.2466447

5. John W. Creswell. 2007. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*.

6. M. Datar, N. Immorlica, Piotr Indyk, and V.S. Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. *Proceedings of the twentieth annual symposium on Computational geometry*: 253–262. http://doi.org/10.1145/997817.997857

7. Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic, and Narayan Bhamidipati. 2015. Hierarchical Neural Language Models for Joint Representation of Streaming Documents and their Content. *International World Wide Web Conference Committee (IW3C2)*.

8. Susan T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* 38: 188–230. http://doi.org/10.1002/aris.1440380105

9. Jacob Eisenstein, Duen Horng Chau, Aniket Kittur, and Eric Xing. 2012. TopicViz: Interactive topic exploration in document collections. *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*, 2177–2182. http://doi.org/10.1145/2212776.2223772

10. Ethan Fast, Binbin Chen, and Michael Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. http://doi.org/10.1145/2858036.2858535

11. Marguerite Fischer. 1966. The KWIC Index Concept : A Retrospective View. April: 57–70.

12. Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJCAI International Joint Conference on Artificial Intelligence*: 1606–1611. http://doi.org/10.1145/2063576.2063865

13. Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*: 929–932. http://doi.org/10.1145/1240624.1240764

14. Marti a. Hearst and Marti a. Hearst. 1998. Automated discovery of wordnet relations.

WordNet: an electronic lexical database: 131–152.

15. Tom Kenter and Maarten De Rijke. 2015. Short Text Similarity with Word Embeddings Categories and Subject Descriptors. *Proceedings of the twenty fourthth ACM International Conference on Information and Knowledge Management ACM International Conference on Information and Knowledge Management*, Vol. 15. 115.

16. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, et al. 2015. Skip-Thought Vectors. *ArxiV*, 786: 1–11.

17. H.P. Luhn. 1960. KEY WORD-IN-CONTEXT INDEX. *American Documentation* XI, 4: 288–295.

18. Eva Martinez and Lluis Marquez. 2014. Document-Level Machine Translation on. *20th International Joint Conference of the European Association for Machine Translation*, 59–66.

19. Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st national conference on Artificial intelligence* 1: 775–780. http://doi.org/10.1.1.65.3690

20. Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*: 233–242. http://doi.org/10.1145/1321440.1321475

21. Thomas Mikolov. 2015. word2vec: Tool for computing continuous distributed representations of words.

22. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Nips*: 1–9. http://doi.org/10.1162/jmlr.2003.3.4-5.951

23. Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*: 1–12. http://doi.org/10.1162/153244303322533223

24. Aditi Muralidharan, MA Hearst, and Christopher Fan. 2013. WordSeer: a knowledge synthesis environment for textual data. *... on information & knowledge ...*: 2533–2536. http://doi.org/10.1145/2505515.2508212

25. Aditi Muralidharan and Marti A Hearst. 2013. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing* 28: 283–295. http://doi.org/10.1093/llc/fqs044

26. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems* 54, 1999–66: 1–17. http://doi.org/10.1.1.31.1768

27. Johnny Saldana. 2009. *The Coding Manual for Qualitative Researchers*. http://doi.org/10.1109/TEST.2002.1041893

28. Helen Marson Smith. 2006. Interpreting Qualitative Data: Methods for Analyzing Talk, Text and Interaction (3rd edition. *Sociological Research Online 11*.

29. Peter D Turney. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning (ECML-2001), Freiburg, Germany*: 491–502. http://doi.org/10.1007/3-540-44795-4_42

30. Yiran Wang, Melissa Niiya, Gloria Mark, Stephanie M. Reich, and Mark Warschauer. 2015. Coming of Age (Digitally): An Ecological View of Social Media Use among College Students. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*: 571–582. http://doi.org/10.1145/2675133.2675271

31. Martin Wattenberg and Fernanda B. Viégas. 2008. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6: 1221–1228. http://doi.org/10.1109/TVCG.2008.172

32. Douglas Zytko, Sukeshini A. Grandhi, and Quentin Jones. 2014. Impression Management Struggles in Online Dating. *Proceedings of the 18th International Conference on Supporting Group Work - GROUP '14*, 53–62. http://doi.org/10.1145/2660398.2660410

33. Wikipedia. Retrieved January 17, 2016 from http://wikipedia.org/

34. 2007. QSR - NVivo Products.

35. 2007. MaxQDA: The art of text analysis.

36. 2007. Atlas.Ti: The Qualitative Data Analysis & Research Software.

37. 2016. hyperRESEARCH.

38. 2016. Sysomos Scout.

39. 2016. HootSuite: The best way to manage social media.

40. 2016. mBlast: Personalized ads and audiences.