Cognitive Assistance in Government
Papers from the AAAI 2015 Fall Symposium

Domain Scoping for Subject Matter Experts

Elham Khabiri, Matthew Riemer, Fenno F. Heath III, Richard Hull
IBM Research
ekhabiri,mdriemer,theath,hull@us.ibm.com

Abstract

Exploring web and in particular social media data is an essential task to many of the subject matter experts in order to discover content around their subject of interest. It is important to provide them with a tool to define their scope of vocabulary, i.e what to search for, and suggest them commonly used terms besides the serendipitous terms allowing them to define their scope of explorations. This paper presents methods on constructing "domain models" which are families of keywords and extractors to enable focus on social media documents relevant to a project using multiple channels of information extraction.

Introduction

News aggregators, forums, blogs, on-line debates, and many other forms of social media are among early successes in the emerging social computing paradigm. Prominent Social Web examples include large-scale information sharing communities (e.g., Wikipedia), social media sites (e.g., YouTube), and web-based social networks (e.g., Facebook), each centered around user-contributed content and community-based information sharing.

There are different levels of content control over each of these resources. a) Some of them are composed of an editorial committee which assure the correctness of the contributed content in the hope of creating a more reliable and well-formed source of information for the public. Wikipedia and editorial selection of comments in NYTimes are among of these cases. b) Other resources, such as tweets, blogs, forum do not let any changes by editorial professionals. The power of such websites comes from huge participation of normal people that shape the crowd intelligence of the web. c) Formal news websites such as CNN and BBC on the other hand, provide daily updates of the news written by professional journalists. The content are again accurate, however it is composed by comparably small number of professionals whose job is dedicated to writing and creating content. We believe all of the three resource types of information that is, formal, semi-formal and informal generated content mentioned above are important to discover relevant concepts and each will shed light on different viewpoint of a target subject. Subject matter experts (SME), such as business owners are willing to understand the demand of the market, what people think about their products and the product of competitors. The strong and weak points reflected by social media will help them to customize and modify their product offerings and assist them to create the most possible relevant responses to the public demands. However discovering relevant content is a significant challenge to many of them. In Government and Public Sector usage our tool could be considered as a cognitive assistant to help discovering relevant concepts to their topic of interest. Such as exploring public opinion about proposed changes in education, insurance and many other relevant topics.

In this paper we propose a domain scoping tool which enables subject matter experts to define their vocabulary of search to find the most relevant content from the social web. We take advantage of many available resources including Google News, Forums, Wikipedia to identify concepts of interest that are relevant to the initial seed terms. Our domain scoping tool advances the state of the art of social media analytics in two fundamental ways. First, the system brings together several text analytics tools to provide a broad-based environment to rapidly create domain models. This contrasts with research that has focused on perfecting such tools in isolation. Second, it applies data-centric and other design principles to provide a working platform that supports ad hoc, iterative, and collaborative exploration of social media data. The domain scoping component presented here is currently part of the Alexandria system (Heath et al. 2015) under development at IBM Research which provides an extensible framework and platform for supporting a variety of big-data analytics and visualizations.

Related Work

Many papers focus on understanding, tagging and ranking of content in social media (Pang and Lee 2008; Khabiri, Hsu, and Caverlee 2009; Fan and Gordon 2014). Various social media studies provide understanding of how information is gathered. For instance, (Leavitt and Clark 2014) analyses community behaviors of social news site in the face of a disaster, (Chew and Eysenbach 2010) studies information sharing on Twitter during bird flu breakout, and (Choudhury, Morris, and White 2014) studies how people use search en-

gines and twitter to gain insights on health information, providing motivation for ad hoc exploration of social data. Fundamentally, the authors of (Rajaraman and Ullman 2011) elaborated on design features needed in a tool for data exploration and analysis, and coined the term "Information Building Applications". They emphasized the support for tagging and categorizing raw data into categorization and the ability to restructure categories as their users, students, understand more about the data or discover new facts. The authors also emphasized the necessity of supporting fluid shift between concrete (raw data) and abstract (category of data) during the validation and iteration process, especially when faced with suspicious outcomes. While the paper discussed specifically about a tool for exploring streams of images, the nature of the approach is very similar to the process of exploring social media we are supporting in our domain scoping tool (DST).

The novelty in our work is the combination of various text analytics and social media exploration tools into a broad-based solution for rapid and iterative domain modeling. While many tools exist, such as Topsy (Topsy Development Team), Solr (SOLR Development Team), Banana (Banana Development Team), we discovered that these tools do not support well the process and the human thoughts in gathering quality results. The existing tools typically tend to aid in a fraction of the overall exploration task needed. More comprehensive, commercial tools such as HelpSocial (HelpSocial Development Team) and IBM Social Media Analytics (SMA Development Team) are geared towards a complete solution. However, these tools require employing a team of consultants with deep domain expertise to operate as consulting services. Their support for the exploration process is not trivial and relies heavily on human labor and expertise.

The domain scoping tool presented here is currently part of the Alexandria system (Heath et al. 2015) under development at IBM Research which provides an extensible framework and platform for supporting a variety of big-data analytics and visualizations. Its architecture is centered around a variety of REST-based service APIs to enable flexible orchestration of the system capabilities; these are especially useful to support knowledge-worker driven iterative exploration of social phenomena. DST is helping to close a key gap in research on tooling for data exploration that was identified in (Bertini and Lalanne 2009). It is focused on enabling rapid exploration of text-based social media data.

Domain Scoping Tool

Domain Scoping addresses the challenge of constructing Domain Models. A Domain Model is typically represented as families of keywords and composite topics (a.k.a., text extractors), which get applied to the corpus of text documents to realize the search or filtering in the corpus. Traditionally, Domain Scoping is performed by a subject matter expert who understands the domain very well and can specify precisely what the particular queries and search criteria should be for a given set of topics of interest. A central goal of Domain Scoping Tool (DST) is to simplify significantly the task of creation of Domain Models as well as to lower the required domain expertise of the person creating Domain Models. To achieve that, we developed several techniques that leverage text analysis and data mining in order to assist at discovery and definition of relevant topics that will drive creation of search queries. In particular, we describe our approach for discovery of relevant collocated terms. This allow very easy, iterative definition of terms and topics (i.e., sets of collocated terms) relevant for a particular domain with minimal input required from the user.

Domain Scoping Tool (DST) includes primarily various analytics on background text corpora that support several functionalities, including similar term generation, parts-of-speech and collocation analytics, and term-frequency-inverse-document-frequency (TF-IDF) analytics and discovery based on using an ontology such as DBPedia.

The system is focused on enabling rapid exploration of text-based social media data. It provides tools to help with constructing "domain models" (i.e., families of keywords and extractors to enable focus on tweets and other social media documents relevant to a project), to rapidly extract and segment the relevant social media and its authors, to apply further analytics (such as finding trends and anomalous terms), and visualizing the results.

Collocated Term Discovery

Domain Scoping Tool (DST), employs two techniques term frequencyinverse document frequency (TF-IDF) score and collocation to discover significant relevant terms to a specific set of seed terms. Simply put, what it does is find documents that seed terms appeared within. This is called the "foreground" documents. It then harvests other terms that were mentioned in the documents and computes their significance. To support this analytic, we acquired sample documents documents considered general and representative enough of many different topics and domains as the "background" materials for this operation. For this purpose we collected a complete week of documents (Sept 1-7 2014) from BoardReader. This extraction amounts to about 9 millions documents. The documents were then indexed in SOLR (SOLR Development Team), a fast indexing and querying engine based on Lucene, for later fast access. Next we queried "NY Times" from this large set of documents, which resulted in news articles in many different areas including politics, sports, science and technology, business, etc. This set of documents is used to build a dictionary of terms that are not limited to a specific domain within a small sample. It is the basis for DST to calculate term frequency in general documents. From the foreground materials, DST computes the significance of other terms in the documents using TF-IDF scores. TF-IDF score is a numerical statistic widely used in information retrieval and text mining to indicate the importance of a term to a document (Manning and Schütze 1999). The score of a term is proportional to the frequency of the term in a document, but is offset by the frequency of the same term in general documents. The TF-IDF score of a word is high if the term has high frequency (in the given document) and a low frequency in the general documents. In other words, if a term appears a lot in a document, it may be worth special attention. However, if the term appears a lot in other documents as well, then its significance

is low.

$$\begin{aligned} TF-IDF &= TF(t,d) \times IDF(t,D) \\ IDF(t,D) &= log\frac{N}{|t \in D|} \end{aligned}$$

To identify relevant phrases to the seed term set, Domain Scoping Tool applies a collocation-based technique which considers the part of speech tagging of the individual terms along with the frequency of appearance of the phrase in the corpus. A collocation is an expression consisting of two or more words that corresponds to some conventional way of saying things. They include noun phrases such as "weapon of mass destruction", phrasal verbs like "make up" and other stock phrases such as "the rich and powerful". We applied collocation to bring in highly relevant terms as phrases when the words collocate in the document and would make no sense as individual terms. More details of this technique can be found in (Bengio and others 2003). For collocated term generation, the larger the corpus and the more accurate the results will be. However a very large corpus will suffer from efficiency and is not practical to use in an interactive environment. Our hypothesis is that a week of general documents as a background corpus is a good enough representative of the bigger corpus, but is small enough to calculate the TF-IDF and collocation scores in a responsive manner.

NNLMs based Similar Term Discovery

Domain Scoping Tool uses Neural Network Language Models (NNLMs), that map words and bodies of text to latent vector spaces. Since they were initially proposed (Bengio and others 2003), a great amount of progress has been made in improving these models to capture many complex types of semantic and syntactic relationships (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). NNLMs are generally trained in an unsupervised manner over a large corpus (greater than 1 billion words) that contains relevant information to downstream classification tasks. Popular classification methods to extract powerful vector spaces from these corpora rely on either maximizing the log-likelihood of a word, given its context words (Mikolov et al. 2013), called *Word2Vec*, or directly training from the probabilistic properties of word co-occurrences (Pennington, Socher, and Manning 2014). In DST, we train our NNLMs on either a large corpus of tweets from Twitter or a large corpus of news documents to reflect the linguistic differences in the target domain the end user is trying to explore. We also extended the basic NNLM architecture to include phrases that are longer than those directly trained in the corpus by introducing language compositionality into our model (Socher and others 2013; Goller and Kuchler 1996; Mikolov et al. 2010). This way, our NNLM models can map any length of text into the same latent vector spaces for comparison.

Co-occurrence based Term Discovery

While the context term discovery approach described above based on targeted documents is useful for generating context words based on very specific concepts, analyzing the broader global scale context of words creates better results for more general concepts. Our approach to global scale context word prediction uses the co-occurrence matrix recently utilized in the context of NNLMs by (Pennington, Socher, and Manning 2014; Levy and Goldberg 2014). Based on the related work this method is called *Glove*. This matrix is created by tracking all co-occurrences within a window size (for example a 10 word window) between N-grams in a predefined vocabulary. N-gram sizes generally range from unigrams to trigrams. The total in context occurrences of word *a* and word *b* is stored in row *a* and column *b* of the co-occurrence matrix. Although this matrix takes up quite a bit of memory, it is manageable and the matrix has no limit in the data quantity it can process to build its semantic understanding as long as the vocabulary is fixed (or at least relatively). Along with the co-occurrence counts with all context words, we also store the total occurrences of each word and maintain a map from N-grams to matrix rows. So, to compute the most probable words to occur in the context of word *j* with an index of *j*, we take the vector representing the *j*th row of the matrix and divide it by the total occurrences of *j*. Next we sort the indexes in order of decreasing probability and map the top indexes to words we can suggest to the user. This approach scales to a group of multiple terms by recursively applying a Hadamard product between the first word's (or current) probability vector and the probability vector of each subsequent word.

DbPedia based Similar Term Discovery

One way of discovering relevant concept to the initial seed terms is to take advantage of ontologies where entities and their relationships are well defined. For example we are able to distinguish entities that belong to the same category. In this paper we leverage the Wikipedia as our ontology resource. Using Wikipedia dumps released by DbPedia, we were able to extract entity-category relationships. For each input entity in the seed term set, DST identifies *peer* entities, meaning the entities that belong to the same category. As an example, a knowledge worker might want to gain insights about relevant concepts to his keyword of search, "Air France". Using Wikipedia, one can see it belongs to multiple categories one of which is "IATA Members". Diving into the category of "IATA Members", many of entities including "Kenya Airways", "Kish Air", "KLM", "Korean Air" will be extracted as relevant concepts. Adding more relevant seed terms such as "KLM", would assist the tool to weigh more on the categories that are common among the input entities. For example, "Association of European Airlines members" will be the common parent of the two seed entities which allows extraction of more similar entities such as "Lufthansa" and "British Airways". DST allows dealing with disambiguations in Wikipedia. It contains an inherent disambiguation module which maps an entity to top k possible relevant entities in Wikipedia using KeywordSearch API (dbpedia-lookup) and allows the SME to select the most relevant disambiguated term. As shown in Figure 8, the term education is mapped to multiple disambiguations including, "Education", "Philosophy of education", "Knowledge sharing" and so on. Each of which belong to one or more categories. The SME could select one or more disambiguations

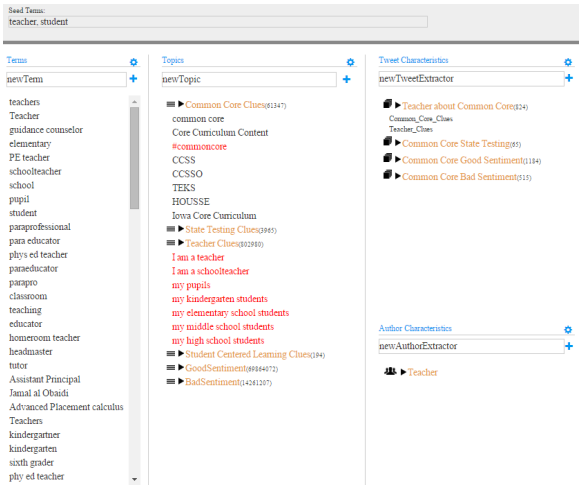

Figure 1: Illustration of Alexandria UI for defining terms, topics, and extractors

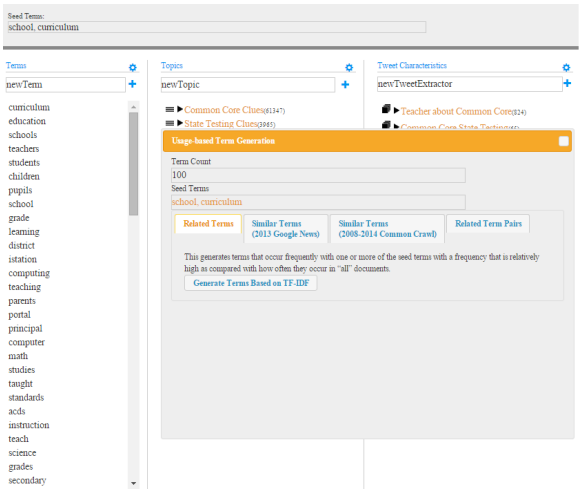

Figure 2: Generation of Related Terms from seed terms "school" and "curriculum", using TFIDF

with one or more category of interest to define his scope of search, in order to retrieve *peer* entities from each selected category.

Domain Scoping in Action

This section provides a brief illustration of how the Domain Scoping Tool can be used in practice. We first illustrate how the Alexandria tool is used to create *extractors* (i.e., queries) against a corpus of text documents (in this example, Tweets), and then illustrate the various term generation capabilities that are supported.

We consider a scenario in which an SME is exploring tweets that about the Common Core State Standards (CCSS), an important education initiative being followed by most of the U.S. states. The SME is interested in topics such as K-12 school curricula, standardized state-level testing, teaching philosophies, public sentiment, etc. The screen shots illustrate different steps that the SME might take during the exploration, and do not reflect a "finalized" family of topics or extractors (please see (Heath et al. 2015).)

Figure 1 shows an interactive, three-column user interface (UI) that is used in Alexandria to create extractors. Users can perform a variety of operations, including dragging and dropping, creating new extractors and topics, and invoking the various term generation routines (illustrated below). As shown in the right column at the top, the extractor Teacher about Common Core is defined as a combination of the Topics Common Core Clues and Teacher Clues. The first of these topics, Common Core Clues, was formed primarily by using terms generated by the DST (see Figure 7 below), although "#commoncore", marked in red, was added after previewing some of the tweets that match against the auto-generated terms. The second topic, Teacher Clues, was created by using terms auto-generated from the seed terms "teacher" and "student", as shown in the left column. However, to create entries in Teacher Clues the auto-generated terms were augmented into phrases such as "I am

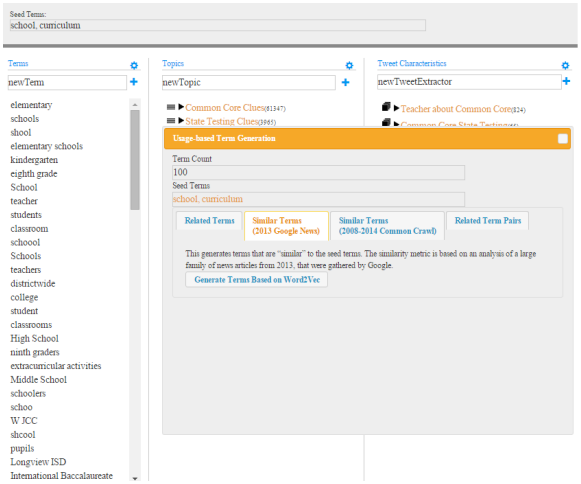

Figure 3: Generation of Related Terms from seed terms "school" and "curriculum", using Word2Vec

a schoolteacher" and "my pupils", to enable a focus on K-12 education rather than college. The extractor Teacher about Common Core will match each Tweet that includes at least one term/phrase from the first topic and one term/phrase from the second topic; the system also supports matching based on arbitrary boolean combinations of topics.

We now illustrate the several mechanisms provided for generation terms in the left column which is the main focus of this paper. Figures 2, 3 and 4 illustrate three of the term generation algorithms, all using the seed terms "school" and "curriculum". The user invokes the operations using the gray pop-up window shown in the figures. The SME might use these activities to start to explore the overall education space, and identify terms that are more specific to K-12 than to university-level education. In this section we quickly illustrate the algorithms; see below for more detail on how they work. Figure 2 shows the output of a TFIDF-

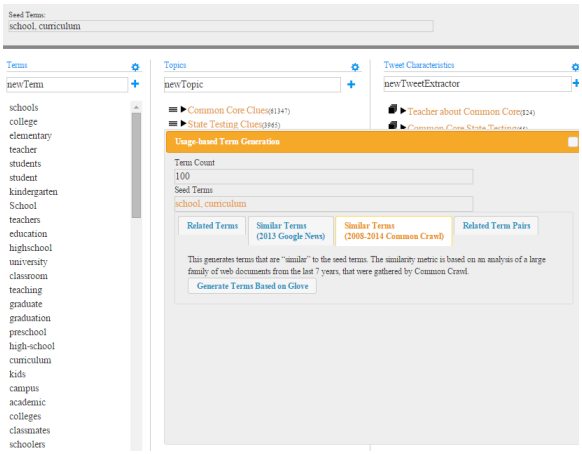

Figure 4: Generation of Related Terms from seed terms "school" and "curriculum", using Glove

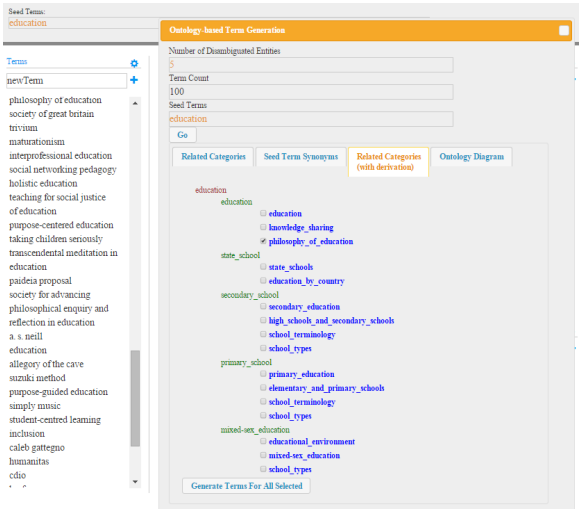

Figure 5: Generation of Categories and selected Terms using DBPedia on "education"

based algorithm, which selects relevant single terms based on the specificity of the terms that are used in the documents that contain the seed terms. Figure 3 shows the output of a Word2Vec-based algorithm that is operating against the 2013 Google News corpus as we explained in the section of *NNLMs based Similar Term Discovery*. Figure 4 show the output of a glove-based algorithm that is operating against the Common Crawl corpus as we explained in the section of *Co-occurrence based Term Discovery.*

The term generation algorithms just described are called *Usage-based*, because they are based directly on corpora of documents that arise naturally in the world. The system also supports *Ontology-based* term generation. For the present the system refers to DBPedia, but other ontologies could also be used (e.g., based on a retailer's product categories or based on a large consulting firm's categorization of industry sectors). Figure 5 shows one application of the DBPedia

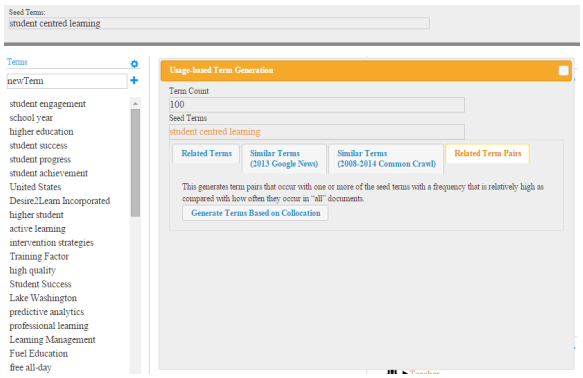

Figure 6: Generation of Related Term Pairs using on seed term "student centred learning", using collocation

ontology, in connection with the seed word "education". As shown here, the system first searches for relevant entities that are included in DBPedia (shown in green), and then finds the DBPedia categories that those entities arise in. For example, "education" and "state school" are entities in DBPedia. The blue entries are the DBPedia categories that contain the entities. In this case, the user has asked to see all of the entities (terms) from DBPedia that arise in the category "philosophy of education". (The user can also request the entities from multiple categories.)

An SME might be familiar with the outputs from DBPedia on education philosophies and/or may refer to Google and elsewhere. For this scenario suppose that the SME wants to explore further the notion of student-centred learning (fifth term from the bottom in Figure 5). Figure 6 illustrates the use of related term generation using a collocated terms algorithm, started with the seed phase "student centred learning". This algorithm is especially useful when searching for bigrams and trigrams that are related to the seed words. The returned phrases include educational goals relating to student centred learning (e.g., student engagement, student achievement), companies in this space (e.g., Desire2Learn), and school districts with an emphasis in this approach (e.g., Lake Washington).

Finally, Figures 7 and 8 illustrate the use of Word2Vec and DBPedia on the seed term "Common Core". The Word2Vec finds terms and entities that are related to the Common Core standard from several different dimensions. This includes state and city level agencies that administer the Common Core standards (e.g., HOUSSE for New York City or Keystone Exams from Pennsylvania), companies and products in this space (e.g., the enVisionMATH Common Core product from Pearson and Math Expressions Common Core from Houghten Mifflin Harcourt), and thought leaders in the Common Core area (e.g., Superintendent Nancy Grasnick).

Figure 8 illustrates the disambiguation feature of the DBPedia-based algorithms. The figure shows that the phrase "Common Core" can refer to the "Common Core State Standards Initiative", which involves education, but can also refer to "Common Core Booseters and Modular Rockets", which arise in the area of rockets and space travel. The trans-

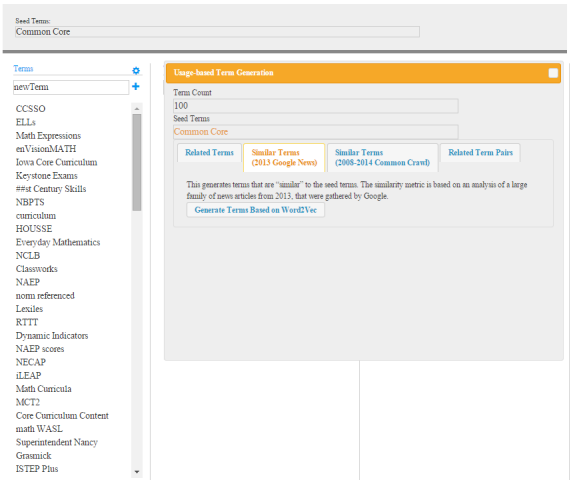

Figure 7: Generation of Similar Terms on seed term "Common Core", using Word2Vec

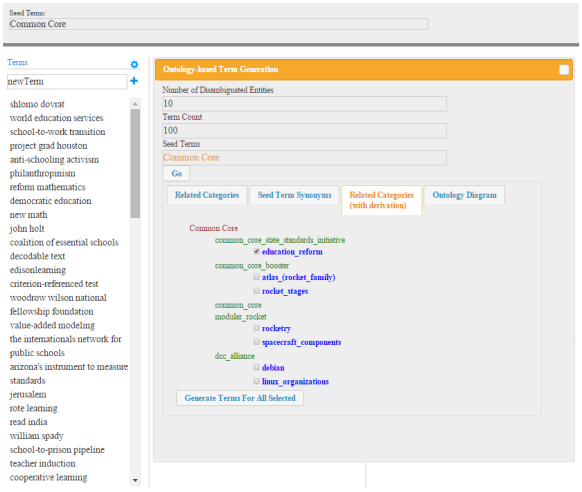

Figure 8: Generation of Categories and Terms using DBPedia on seed term "Common Core"

parency into the DBPedia entities and categories found helps the user target the information of interest.

Conclusion

We presented a Domain Scoping Tool (DST) that facilitates the task of creation of Domain Models by lowering the required domain expertise of the person creating Domain Models. We developed several techniques that leverage text analysis and data mining in order to assist discovery of relevant topics that will drive creation of search queries. DST operates on both Usage-based and Ontology based methods, because it both based on corpora of documents that arise naturally in the world as well as using Ontologies with well-defined relationships. This allow very easy, iterative definition of terms and topics (i.e., sets of collocated terms) relevant for a particular domain with minimal input required from the user.

References

Banana Development Team. Banana. https://docs.lucidworks.com/display/SiLK/Banana.

Bengio, Y., et al. 2003. A neural probabilistic language model. *J. of Machine Learning Research* 3:1137–1155.

Bertini, E., and Lalanne, D. 2009. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Explorations* 11(2):9–18.

Chew, C., and Eysenbach, G. 2010. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLOS ONE* 5(11). e14118. doi:10.1371/journal.pone.0014118.

Choudhury, M.; Morris, M.; and White, R. 2014. Seeking and sharing health information online: Comparing search engines and social media. In *Proc. ACM Intl. Conf. Computer Human Interaction (CHI)*, 1365–1375.

dbpedia-lookup. Dbpedia-Lookup. http://wiki.dbpedia.org/projects/dbpedia-lookup.

Fan, W., and Gordon, M. D. 2014. The power of social media analytics. *Communications of the ACM* 57(6):74–81.

Goller, C., and Kuchler, A. 1996. Learning task-dependent distributed structure-representations by backpropagation through structure. *IEEE International Conference on Neural Networks* 347–352.

Heath, F. F.; Hull, R.; Khabiri, E.; Riemer, M.; Sukaviriya, N.; and Vaculin, R. 2015. Alexandria: Extensible framework for rapid exploration of social media. In *Proceeding of IEEE Big Data Congress*. IEEE.

HelpSocial Development Team. HelpSocial. https://www.helpsocial.com/.

Khabiri, E.; Hsu, C.-F.; and Caverlee, J. 2009. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community. In *3rd Intl AAAI Conference on Weblogs and Social Media ICWSM*, volume 9.

Leavitt, A., and Clark, J. 2014. Upvoting Hurricane Sandy: Event-based new production processes on a social news site. In *Proc. ACM Intl. Conf. Computer Human Interaction (CHI)*, 1495–1504.

Levy, O., and Goldberg, Y. 2014. Linguistic regularities in sparse and explicit word representations. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Baltimore, Maryland, USA, June. Association for Computational Linguistics*.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Proc. 11th Ann. Conf. of the Intl. Speech Communication Association (INTERSPEECH)*, 1045–1048.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1-2):1–135.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proc. 2014 Conf. Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar.*, 1532–1543. ACL.

Rajaraman, A., and Ullman, J. D. 2011. *Mining of massive datasets*. Cambridge University Press.

SMA Development Team. IBM Social Media Analytics. http://www-01.ibm.com/software/analytics/solutions/customer-analytics/social -media-analytics/.

Socher, R., et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, 1642. Citeseer.

SOLR Development Team. SOLR Home Page. http://lucene.apache.org/solr/.

Topsy Development Team. Topsy. http://topsy.com.