Unsupervised Text Classification and Search using Word Embeddings on a Self-Organizing Map

Suraj Subramanian
Vidyalankar Institute of Technology
University of Mumbai
Mumbai, India

Deepali Vora
Asst. Professor
Vidyalankar Institute of Technology
Mumbai, India

ABSTRACT

This paper presents the results of an experimental implementation of a document classifier leveraging contextual word embeddings clustered on a self-organizing map. The problem of document categorization is further compounded when there are no predefined categories, or conversely there are too many categories, that documents may be bucketed into. This paper proposes to address these problems by modelling the major themes contained in the document corpus into a cluster-map using a self-organizing neural network. The cluster-map provides a visual representation to explore the corpus, and a near-semantic search interface of the many concepts outlined across the corpus.

Keywords

Clustering, knowledge retrieval, natural language processing, neural nets, self organizing map, topic modelling, semantic search, unsupervised.

1. INTRODUCTION

Supervised text categorization is used to segregate a corpus of documents into classes that are predefined. This is achieved by training the model on a number of pre-classified examples, and a large corpus of previous work on this exists. One problem with this approach is that it is difficult and time-consuming to label training data manually. Classification without predefined target categories addresses this problem by clustering related documents solely on the basis of their contents.

Finding a classification method that can be applied in concept to virtually any ecosystem (size/language/domain) of document corpora is the main challenge of this project. Another equally complex challenge is finding the optimal classes that the documents can be bucketed into. Since there are a huge number of individual concepts in any given corpus, it is vital to select only those that have a sizable presence in the corpus.

This paper explores a method that represents any corpus of documents (independent of language or domain) in vector space and groups them into categories most relevant to their contents using a Kohonen Self Organizing Map, while generates a visually intuitive representation of the corpus in relation to the user's search queries.

2. RELATED WORKS

Text mining combines the disciplines of data mining, information extraction, information retrieval, text categorization, machine-learning, and computational linguistics to discover structure, patterns, and knowledge in large textual corpora. Text mining by using self-organizing map (SOM) techniques had gained a spurt in attention in knowledge discovery research and the information retrieval field. A paper by Lin et al. [7] perhaps marks the first attempt to utilize SOM (unsupervised neural networks) for information retrieval work.

In addition, this paper certainly must mention WEBSOM [1],[2],[3]. The goal of this seminal work is the exploration of document collections by topic. The WEBSOM uses the self-organizing semantic map [4] to identify context of words; this vocabulary representation is used in the preprocessing stage to encode documents. These map-encoded documents are then automatically clustered according to where their contained words lie on the vocabulary map. Documents containing similar terms and words are thus clustered closer to each other; this feature facilitates a visually intuitive exploration of the corpus; users can zoom in on groups of documents related to a very specific group of words.

This project is heavily influenced by the findings of WEBSOM. Differing from WEBSOM, this paper does not employ the use of a self-organizing semantic map; instead it employs the use of CNN-trained word embeddings to capture contextual information. In addition to the visual representation of the corpus, concept-based querying for document retrieval is provided. The system allows a user to input a keyword similar to a search engine. This returns a list of suggested documents that a) lie in a cluster most relevant to the keyword, and b) contain concepts that are significantly relevant to the keyword, regardless of the actual presence of that keyword in the document. The described technique is also a language-neutral method with the only requirement being a sizable corpus to learn the language's vocabulary.

2.1 Contextual Word Vectors

This paper utilizes the contextual information in continuous vector representations of words from large a large corpus of text (Mikolov et al., 2013c). The statistics of word occurrence are used to model the probability of co-occurrence of the words in the neighbourhood of a given focus word I [5] or the ratios of this co-occurrence probability to the probability of I's occurrence (Pennington et al., 2014) [6].

GloVe is a count-based model, as opposed to the predictive model that word2vec is. It essentially constructs a large combinatorial matrix of (word x context) which is then normalised, log-smoothed and factorised to reduce dimensionality.

The method described in this paper relies on the contextual relationships between words, which is captured in the GloVe model by virtue of operating directly on the co-occurrence statistics of the corpus.

3. DATA PREPROCESSING

This section describes in detail the steps taken to prepare the vector space representation of the documents to be clustered on the SOM. Data preprocessing is an essential and critical step for effective clustering. The first part is to process the document lexicon, which involves stopword removal and part-of-speech tagging. The second preprocessing step is to map the minimized document into vector space.

3.1 Stopword Removal

The first step is to obtain a list of words that satisfactorily represent a document. Words like conjunctions, interjections are relevant for syntactic understanding of the text but do not possess any information about the topic of the document. To minimize this noise, we filter out stopwords (from NLTK), punctuation, digits and non-words.

Interest in documents often centers on entities which is typically represented in the document's nouns [8]. Though in some cases, it is important to consider adjectives and verbs (eg: sentiment analysis), the scope of this project is to identify topics qualitatively; hence a noun-only approach is followed here.

3.2 Vector Space Mapping

This paper uses the word vectors described in section 1.1 to map all the preprocessed documents in the vector space. Each word is mapped to a vector from the pre-trained vector space, and stored along with its frequency in the document.

The order of occurrence of the phrases in the documents is retained for context. The word vectors in each document also undergo k-means clustering, and the centroids of all clusters are stored in a list. These will be later used when searching the corpus.

Finally, the word-vectors, their count-weights and centroid-vectors are stored in an object pointing to the original document.

4. DOCUMENT CLUSTERING

This section details the map training methodology and the algorithm used. 2 passes are made over the document corpus – training (Section 4.1) and classifying (Section 4.2). The training stage adjusts the map's weight vectors to represent closely related words. The classification stage assigns each document to a set of closely related nodes. The search methodology (Section 4.3) enables an end user to explore the corpus visually and retrieve documents that are semantically significant to the search query

4.1 SOM Training

The vocabulary cluster map that is employed to distribute the documents in vector space is produced by measuring the similarity of the words to each other. Words that are related to each other in concept will have a shorter cosine distance in the word-vector space. This property allows co-related words to fall into the same or neighboring map nodes. By means of the SOM algorithm, word or "topic" clusters can be ordered and organized as nodes on the map.

Let

$$\mathbf{v_i} \in \mathbf{R}^{(J\times N)}$$

be the list of word vectors of the ith document in the corpus, where N is the number of features comprising the word vector, and J is the number of words in the document.

Let

$$\mathbf{x_i} = (1, 1, \ldots, 1)\mathbf{v_i} \in \mathbf{R}^N$$

be the document-vector of the ith document, which is derived by summing up the constituent word vectors. These vectors are used as the training inputs to the map. The map is nothing but a grid of processing units called neurons. Each neuron in the map has N "synapses", or a randomly-initialized N-dimensional weight vector.

Let

$$\mathbf{W} = \{\mathbf{w_k} \mid 1 \le k \le K\} \in \mathbf{R}^{(K\times N)}$$

be the synaptic weight vectors of the neurons in the map, where K is the number of neurons on the map. The map is trained by the SOM algorithm:

1. Randomly select a training vector $\mathbf{x_i}$ from $\mathbf{X}$.
2. Find the best matching neuron k with synaptic weights closest to $\mathbf{x_i}$, i.e.

   $$\mathbf{x_i}-\mathbf{w_k} = \mathrm{argmin}_c\, \mathbf{x_i}-\mathbf{w_c}$$

3. For every neuron l in the neighbor of node k, update its synaptic weights by

   $$\mathbf{w_l} \mathrel{+}= \alpha(t)(\mathbf{x_i}-\mathbf{w_l})$$

   where α(t) is the learning rate at time t.
4. Increase time t. If t reaches the preset maximum training time T, stop the training process; otherwise decrease α(t) and the neighborhood radius, and go to Step 1.

After all the iterations, the result is a topological clustering of word vectors, with each neuron's weight representing an amalgamation of related words

4.2 Document Classification

Once the map is obtained, each of the documents in the collection can be classified to the best matching neuron. The k-means centroids obtained for each document are representations of its core concepts. A relevance factor to each neuron is computed using weighted means, where the weights are the size of the centroid cluster. Concurrently each neuron, hitherto represented only by its weight, is assigned keywords of its concept. This serves as a visual aid for exploration of the corpus, to identify which topics can be located in a given neuron.

4.3 Document Searching

The neuron with weight vector closest to the search query vector is deemed to have the most relevant results. The results may be further refined by comparing the search query to the k-means centroids of each document returned in the resultset. A visual representation of the corpus, with respect to the search term, is available to the user. Each neuron has a number ranging from 0 – 255 denoting the relevance factor to the search query. Additionally, neurons deemed to be closer in concept to the search term are shaded brighter, while unrelated neurons are a dull gray colour. This provides an intuitive understanding of the various topics in the corpus that are related to the user's search query.

Once the nearest cluster is identified, focus is narrowed to the documents bucketed within this cluster.

5. IMPLEMENTATION & RESULTS

The word embeddings were obtained from the GloVe website. For this project we used 55-dimensional vectors pretrained on the English Wikipedia corpus.

A 15×15 map was trained on a collection of the NSF Grant Awards 2014. This corpus was chosen for the sheer range in topics and multiple possible class hierarchies inherent to the corpus. The corpus was cleaned of headers to retain only the body content and preprocessed as per the steps mentioned in Section 3. Subsequently the map was trained as per the approach detailed in Section 4.

A search for "education improve" on the NSF Grant Awards 2014 collection was queried by the user. Table 1 contains the labels of the nodes which are deemed to be most relevant to "education improve"

Table 1. BMU attributes for "education improve"

Coordinates	Labels	Score
(1, 11)	Activities, outreach, opportunities, institutions, internships, organizations, educators, workshops, programs, conferences	255
(2, 11)	Expertise, research, contributions, interests, innovations, methodology, innovation, scientists, knowledge, specialists	240
(1, 10)	Partnerships, organizations, stakeholders, stakeholder, partners, entrepreneurs, organization, policymakers, consultants, corporates	223
(2, 10)	Development, resource, sustainability, infrastructure, planning, resources, management, innovation, environment, governance	217

6. CONCLUSION

This paper demonstrates that document clustering can be an effective information access tool in its own right. It is particularly helpful in situations in which it is difficult or undesirable to specify a query formally. To support fast search capabilities, the self-organizing map implemented in this paper is an extremely simple and primitive stateful neural net. However it's efficacy in clustering data points spatially provides a handy visual exploration tool. Coupled with contextual vectors that are essentially the learned vocabulary mapped in hyperspace, we are able to reduce dimensionality sufficiently to allow fast classification and search.

Since it does not require formal ontologies and lexical hierarchies specific to a language, using it for other languages is only a matter of training a contextual word vector model with sufficient text of the intended language.

Improvisations to this project can be in the form of support for stream analysis, where the model actively learns an online stream of text, thereby categorizing documents on the fly.

7. REFERENCES

[1] Honkela, T., Kaski, S., Lagus, K. and Kohonen, T., "Newsgroup exploration with WEBSOM method and browsing interface, " Technical report, vol. 32, 1996.

[2] Kohonen, T., "Self-organization of very large document collections: State of the art," Springer London ICANN 98, pp. 65-74, 1998.

[3] Kaski, S., Honkela, T., Lagus, K. and Kohonen, T., "WEBSOM–self-organizing maps of document collections," Neurocomputing 21(1), pp.101-117, 1998

[4] Ritter, H. and Kohonen, T., "Self-organizing semantic maps," Biological Cybernetics, 61(4), pp.241-254, 1989.

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., "Distributed representations of words and phrases and their compositionality," Advances in Neural Information Processing Systems (pp. 3111-3119), 2013.

[6] Pennington, J., Socher, R. and Manning, C.D., "Glove: Global Vectors for Word Representation," EMNLP, vol. 14, pp. 1532-43, October 2014.

[7] Lin, X., Soergel, D. and Marchionini, G., "A self-organizing semantic map for information retrieval," Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 262-269, September 1991.

[8] Martin, F. and Johnson, M., "More Efficient Topic Modelling Through a Noun Only Approach," Australasian Language Technology Association Workshop, pp. 111, 2015.