Feature-Rich Encoding for Uplift Modelling

Chapman Siu

Abstract

There have been many approaches for word embeddings which have been explored including `word2vec` and `tfidf`.

In this paper we will explore whether enriching our data using approaches using linguistic approaches such as POS and NER tagging will provide substatial uplift to our models which was described as "Feature-Rich Encoding"(Nallapati, Xiang, and Zhou 2016).

Keyword - Text Mining

Introduction

Text document classification is a task of classifying a document into predefined categories based on the contents of the document. A document is represented by a piece of text expressed as words or phrases. The task of traditional text categorization methods is done by human experts. It usually needs a large amount of time to deal with the lack of text categorization. In recent years, text categorization has become an important research topic in machine learning and information retrieval. It has also become an important research topic in text mining, which analyses and extracts useful information from texts. More Learning techniques has been in research for dealing with text categorization.

In recent years there have been lots of interest in various approaches for word embeddings, which generally fall into the following categories:

- term frequency approaches, including term frequency-inverse document frequency approaches (TFIDF), and latent semantic analysis approaches
- word2vec approaches including skip-gram and continuous bag of words approaches (cbow)
- latent dirichlet allocation

Preliminaries

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a well-known topic modelling technique proposed by (Blei, Ng, and Jordan 2003). LDA is a generative probabilistic model of a textual corpus (i.e., a set of textual documents), which takes a training textual corpus as input, and a number of parameters including the number of topics (K) considered. In the training phase, for each document s, LDA will compute its topic distribution θ_s, which is a vector with K elements, and each element corresponds to a topic. The value of each element in θ_s is a real number from 0 to 1, which represents the proportion of the words in s that belong to the corresponding topic in s. After training, LDA can be used to predict the topic distribution θ_m of a new document m. In our case, a document is the description of a question, and the topic is a higher level concept corresponding to a distribution of words. For example, we may have the topic "admissions", which is a distribution of words such as "citizenship", "GRE", "TOEFL", "transcripts".

Word2Vec

Word2Vec is all about computing distributed vector representations of words. In this project we will be using the skip-gram variant.

The training objective of skip-gram is to learn word vector representations that are good at predicting its context in the same sentence. Mathematically, given a sequence of training words $w_1, w_2, ..., w_T$, the objective of the skip-gram model is to maximize the average log-likelihood

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{j=k}\log \Pr(w_{t+j}|w_t)$$

where k is the size of the training window.

In the skip-gram model, every word w is associated with two vectors u_w and v_w which are vector representations of w as word and context respectively. The probability of correctly predicting word w_i given word w_j is determined by the softmax model, which is

$$\Pr(w_i|w_j) = \frac{\exp(u_{w_i}^T v_{w_j})}{\sum_{l=1}^{V}\exp(u_l^T v_{w_j}}$$

where V is the vocabulary size.

The skip-gram model with softmax is expensive because the cost of computing $\log(\Pr(w_i|w_j))$ is proportional to V, which can be easily in order of millions.

Latent Semantic Indexing

Latent Semantic indexing is a transformation on bag-of-words models by applying truncated SVD to term-document matrices. This can be performed on word counts or tf-idf (term frequency-inverse document frequency).

Proposed Approach

In this section we will present the overall framework for our feature-rich encoder. We will consider the original one proposed by (Nallapati, Xiang, and Zhou 2016) and extending it with other commonly used word embeddings such as LDA.

Overall Framework

The framework for this feature building consists of building the relevant word representations. It consists of:

- word2vec
- POS, with word2vec embedding
- NER, with word2vec embedding
- TFIDF
- LDA

In order to build these features, various Python libraries were used, including `scikit-learn`(Buitinck et al. 2013) in order to combine all the features together, `gensim`(Rehurek and Sojka 2010) for the LDA and word2vec implementations, and `nltk`(Bird, Klein, and Loper 2009) which is used to extract POS and NER information.

Supervised Models

In order to compare uplift in performance, we will compare the uplift gained from various popular models which are used in text classification and machine learning including naive bayes(Aghila and others 2010), svm(Joachims 1998), decision trees(Witschel, n.d.).

Experiments and Results

Experimental Setup

The data sets used for this project was the 20 new groups dataset(Joachims 1996) and various datasets from UCI

Evaluation Metrics

Accuracy was used as inline with results discussed in (Joachims 1996).

References

Aghila, G, and others. 2010. "A Survey of Na\" Ive Bayes Machine Learning Approach in Text Document Classification." *ArXiv Preprint ArXiv:1003.1795*.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. " O'Reilly Media, Inc."

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, et al. 2013. "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project." In *ECML Pkdd Workshop: Languages for Data Mining and Machine Learning*, 108–22.

Joachims, Thorsten. 1996. "A Probabilistic Analysis of the Rocchio Algorithm with Tfidf for Text Categorization." DTIC Document.

———. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In *European Conference on Machine Learning*, 137–42. Springer.

Nallapati, Ramesh, Bing Xiang, and Bowen Zhou. 2016. "Sequence-to-Sequence Rnns for Text Summarization." *ArXiv Preprint ArXiv:1602.06023*.

Rehurek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Witschel, Hans Friedrich. n.d. "Using Decision Trees and Text Mining Techniques for Extending Taxonomies." In. Citeseer.