Semantic Annotation of Documents using wikipedia topics

Team 20

Problem Statement

Assigning wikipedia topics to any document using

- Support Vector Machines
- Logistic Regression
- Random forests
- KNN
- or any popular classification algorithm.

Methodology

Data Pre - Processing.

Training word and paragraph vectors.

Training new vectors.

Training the classifier using supervised learning.

Testing the classifier

Pre - Processing

Converting the data into train, test texts.

Training word and paragraph vectors

Word2vec representation is used to train words in the corpus

Doc2vec representation is used to train paragraphs in the corpus

Dimension, which can be tweaked, is set to 400

Epochs can be set to 50 for deep learning

All the two parts of the corpus are represented in the vector space model.

Training and testing the classifier

Classifier takes list of arrays corresponding to its labels.

Training is done on this data.

The testing data sent to the classifier and the accuracies are noted.

Corpus Used

http://www.cs.cornell.edu/people/pabo/movie-review-data/ - Movie Review data

Training data - 25000

- Positive: 12500
- Negative: 12500

Testing data - 12500

Feature Selection

The dimensions of the vector representation of the paragraph is taken as the features of the data and trained.

It is up to our novelty to set the no of features.

More the number of features, the more accuracies.

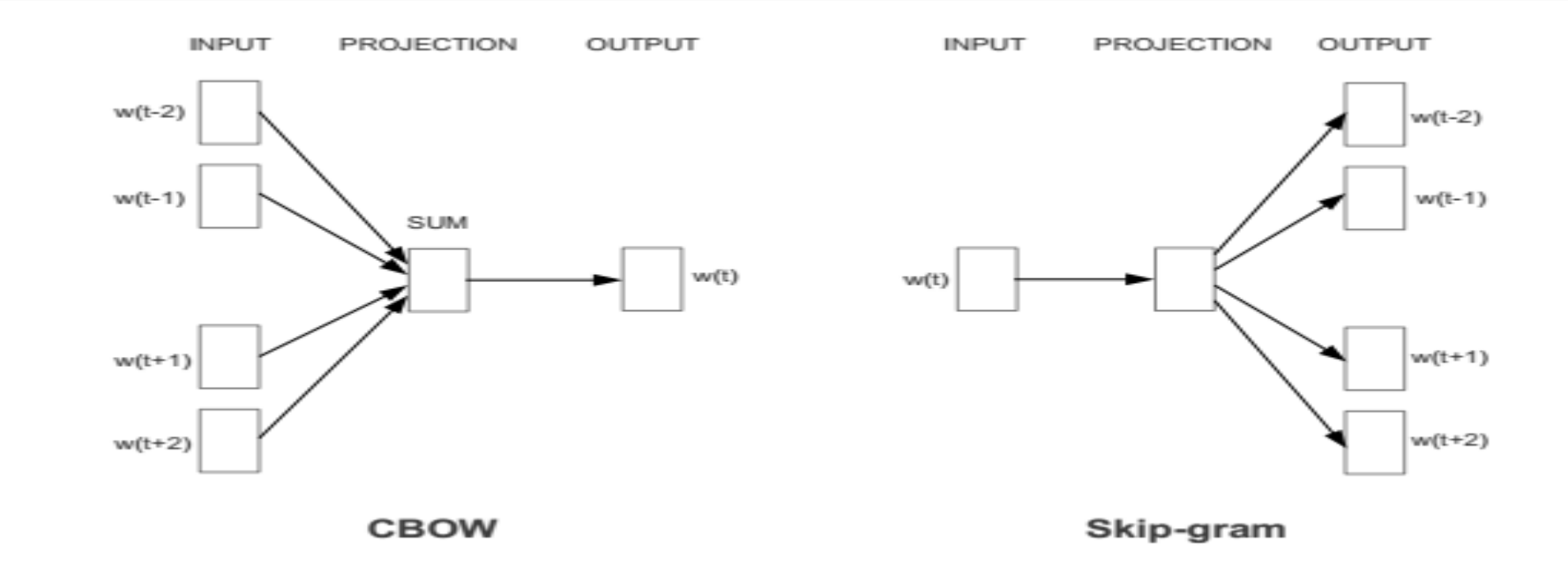
Architectures

As stated in the paper, there are 2 architectures, continuous bag of words based (CBOW) and the other skip-gram (PV-DBOW) based

word2vec and doc2vec are trained using individual and both architectures and the results are visualised.

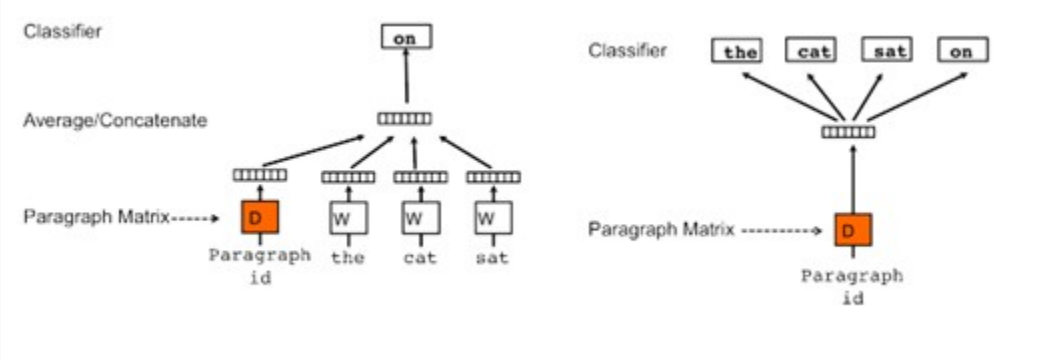
Doc2vec is almost identical to Word2Vec, except we now generalize the method by adding a paragraph/document vector.

Architecture for CBOW and Skip-gram method

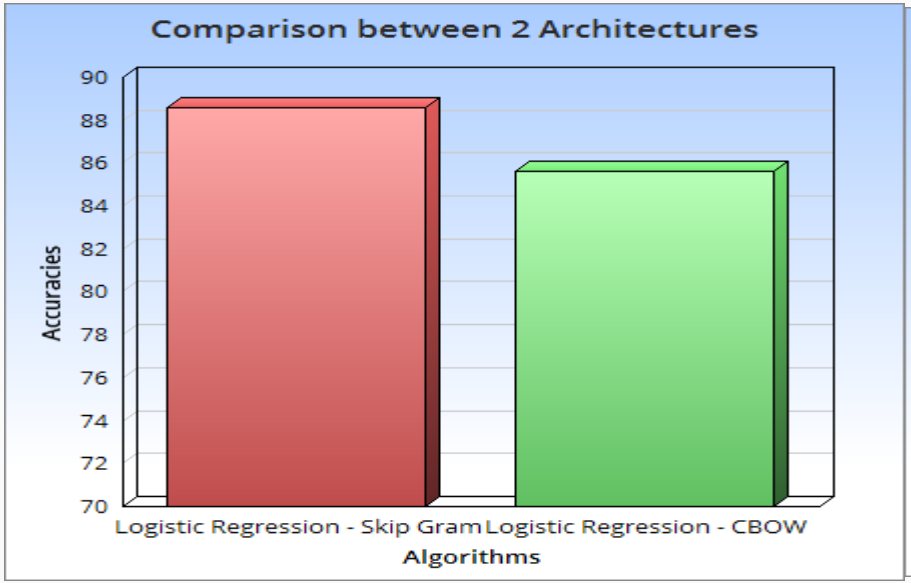


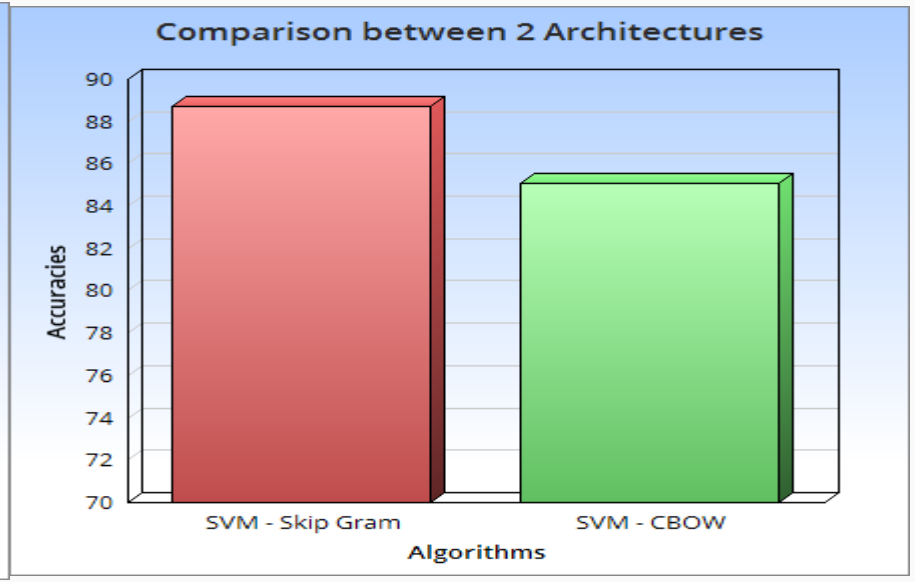
Architecture for Doc2vec

- DM(Distributed Memory) attempts to predict a word given its previous words and a paragraph vector. Even though the context window moves across the text, the paragraph vector does not (hence distributed memory) and allows for some word-order to be captured.
- DBOW (Distributed Bag of Words) predicts a random group of words in a paragraph given only its paragraph vector



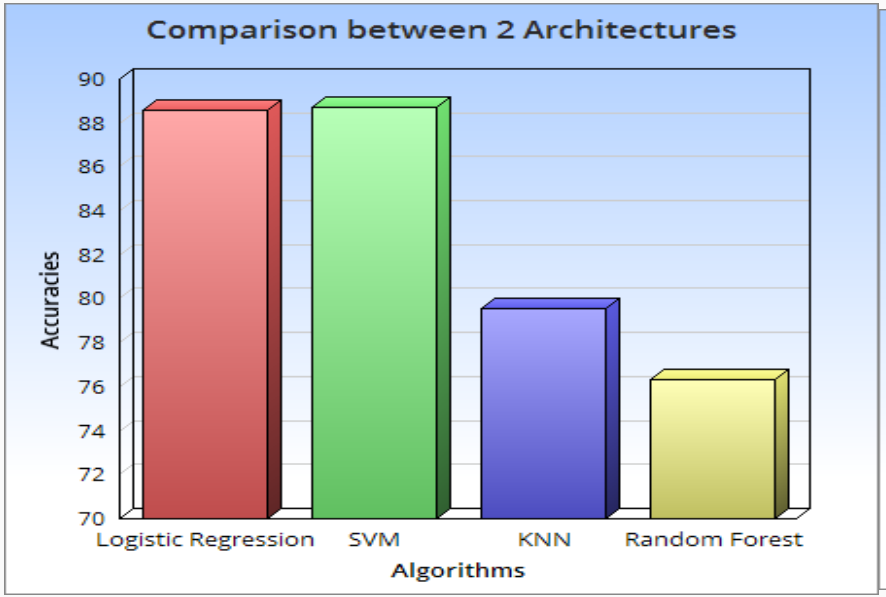
Graph showing comparison between two architectures - logistic regression and SVM



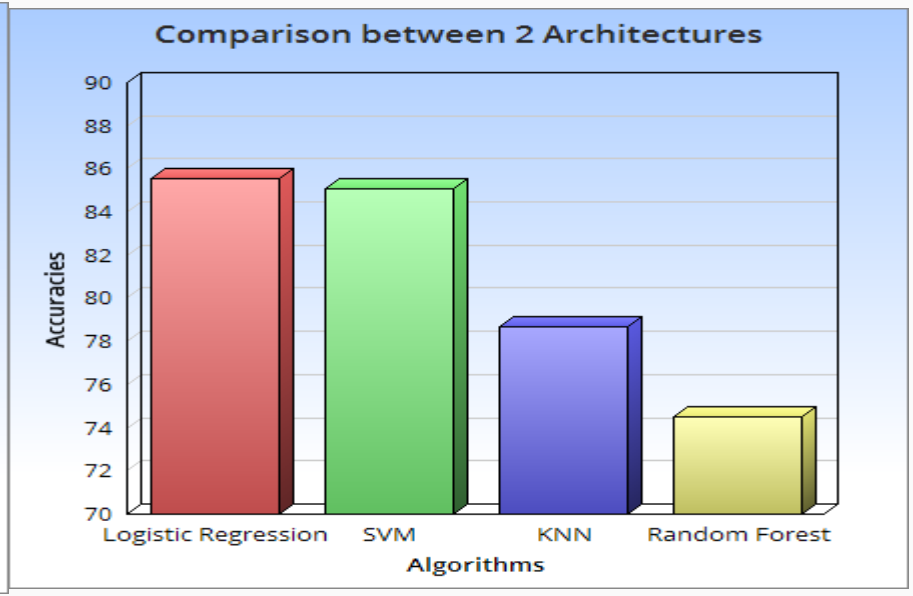


… so on for other graphs also

Graph showing comparison between classification algorithms for both Arch's



Skip Gram



CBOW

Next Step

As said by the mentor, we will be implementing this taking ACM classification tree as the categories and the research papers with abstracts.

Concatenating CBOW and Skip-gram vectors for classification

References

http://rare-technologies.com/doc2vec-tutorial/

https://github.com/piskvorky/gensim

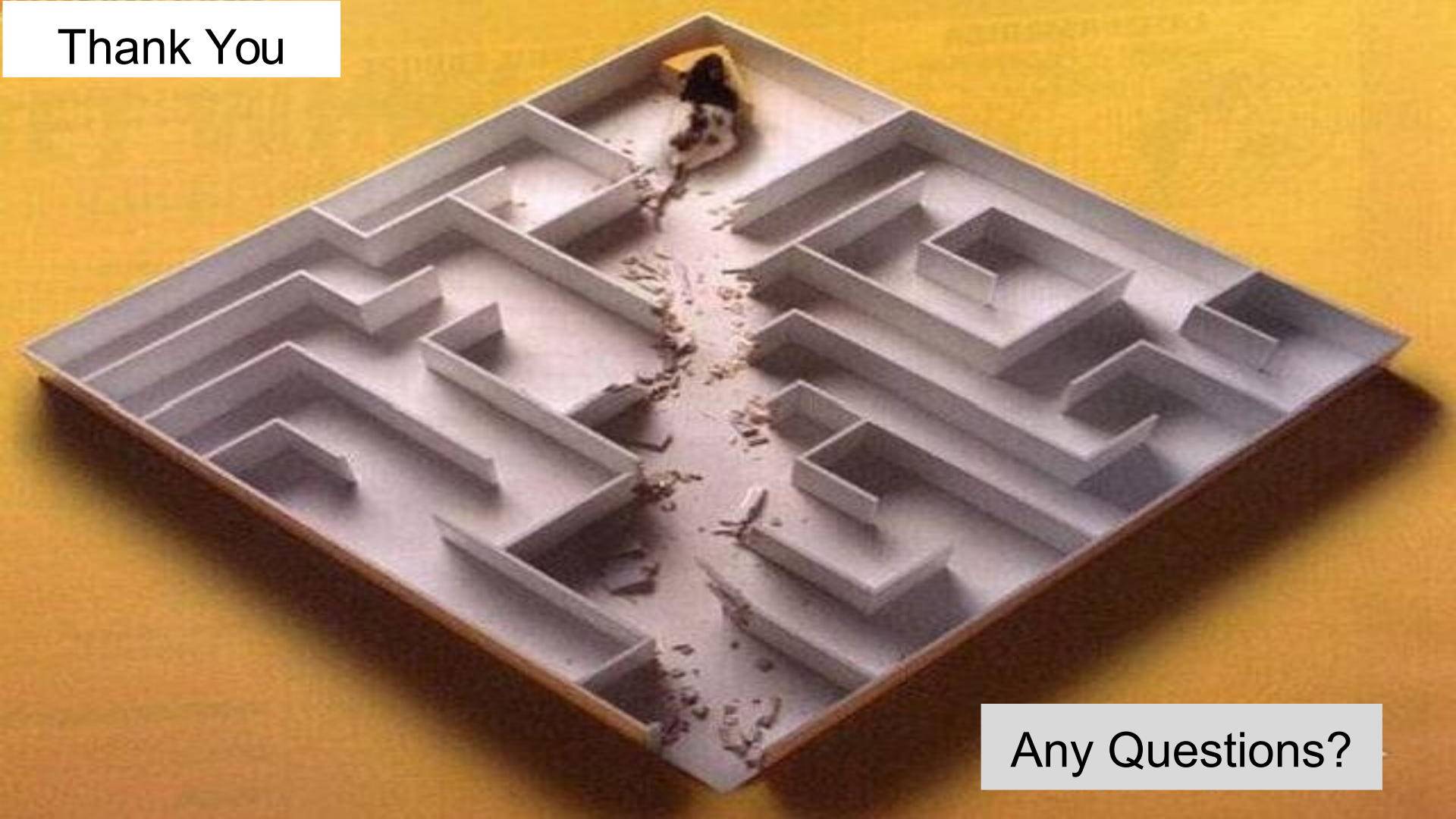
Quoc V. Le, and Tomas Mikolov, ``Distributed Representations of Sentences and Documents ICML", 2014

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, ``Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR", 2013.

Github Link

https://github.com/rohitsakala/semanticAnnotationAcmCategories

Thank You



Any Questions?