**Group No - 11**

**Project No – 2**

**Project Title- Set Expansion**

**Mentor- Nikhil Priyatam Pattisapu Arsh Anand**

**Akshita Negi**

**Ankur Agarwal**

**Achira Maloo**

***Abstract***

As the Internet keeps growing, the type of information available through the web is very diverse. Hence there is a need to develop methods that take advantage of the huge amount of information on the web.

Our project focuses on finding the sets of related words using the World Wide Web.

Set Expansion is the task of finding a more complete Named Entity set from a set of a small number of Named Entities in same type. In other words Set Expansion is the problem of finding all instances of a set given a small number of “seed” instances.

For example- Given names of some countries, set expansion finds the name of other countries.

In this project we will restrict ourselves to the domain of computer science courses only.

***Introduction***

The main objective of the project is to find the set of related words given a few words as input which are related to each other in some or the other manner. To perform the set expansion we make use of the word2vec model provided by the deeplearning4j. Word2Vec is a neural net that processes text before that text is handled by deep-learning algorithms. While it does not implement deep learning, Word2Vec turns text into a numerical form that deep-learning nets can understand- the Vector. Word2Vec creates features without human intervention including the context of individual words. Given enough data, usage and context, Word2vec can make highly accurate guess about a word’s meaning bases on its past appearances. The output of the word2vec neural net is a vocabulary with a vector attached to it, which can be fed into a deep-learning net for classification or labelling. In our project we have used the word2vec model to calculate the similarity between the two words. Word2vec provides this similarity function which returns the cosine similarity between any given two words. Based on the values obtained by the similarity function we expand our given set of related words. We stop this process until we get ten terms which are related to each other. However any other criteria could also be used to stop the process.

***Related Work***

There is a significant body of related work in the broad space of information extraction and named entity extraction. Following are a few systems relevant to set expansion-

**SEAL** System – Wang and Cohen developed the SEAL system for set expansion using a two-phase extraction/ranking architecture.

**Snowball** System – Agichtein introduced this system that bootstraps from a small number of input tuples, by first obtaining typical contextual patterns of the seed tuples from the web pages, which are used in turn to extract more tuples.

**KnowItAll** System – Etzioni developed the KnowItAll system that automatically extracts facts from the web using textual patterns like “cities such as Paris, London and New York” to extract candidate entities. Candidates are then ranked in a bootstrapping manner using statistical information gathered from the search engine.

**Google Sets** – Google sets is a simple tool. It takes as input three or more items that were somehow related to each other and Google attempted to find more members of the set.

***Approach***

First and the most important step is to train the **word2vec** model from a dataset which contains information related to the computer science domain. We obtained such a dataset from the **dmoz** corpus. After training the word2vec model, set expansion technique is applied. The input is given as a file containing the initial seed list or the words which are related to each other in some or the other way. Next step is building queries for the search engine. Any of the search engine like google, bing, faroo, yandex, twitter etc can be used for this purpose. Next we retrieve the top results obtained after querying the search engine. The contents of these retrieved webpages are then parsed and tokenized to obtain different terms present in these webpages. These terms thus obtained are then compared with the current seed terms using the similarity function of the word2vec model which is trained beforehand. The term having the maximum similarity with the current seed terms is considered to be the new seed and is assumed to be related to original set of words. In this way the process continues until we have a set of ten related words. This can be easily extended to find more than ten related words. Thus we get as output the set of related words.

***Conclusion and Future Work***

The given set of related words is used to obtain more words which are related to them. The current project is restricted to the computer science domain. This can be extended to a broad domain with the help of larger dataset. The word2vec model can be trained on this large dataset to perform set expansion on various other domains. Advance techniques can be used like assigning weights to words which occur in some lists, references etc to improve the efficiency of the set expansion technique.

***References***

A cross-lingual dictionary for English Wikipedia concepts

Automatic named entity set expansion using semantic rules and wrappers for unary relations

Automatic set instance extraction using the web

Entity list completion using set expansion techniques

Language-independent set expansion of named entities using the web

SEISA: Set Expansion by Iterative Similarity Aggregation

Web-Scale Distributional Similarity and Entity Set Expansion

Identifying sets of related words from the world wide web

<http://deeplearning4j.org/word2vec.html>

<http://rdf.dmoz.org/rdf/>