A Word Lexicon Builder Using Neural Word Embeddings

Dáire Murphy

Supervisor: Derek Greene

A thesis presentation presented in part fulfilment of the degree of
BSc. (Hons.) in Computer Science with Data Science

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

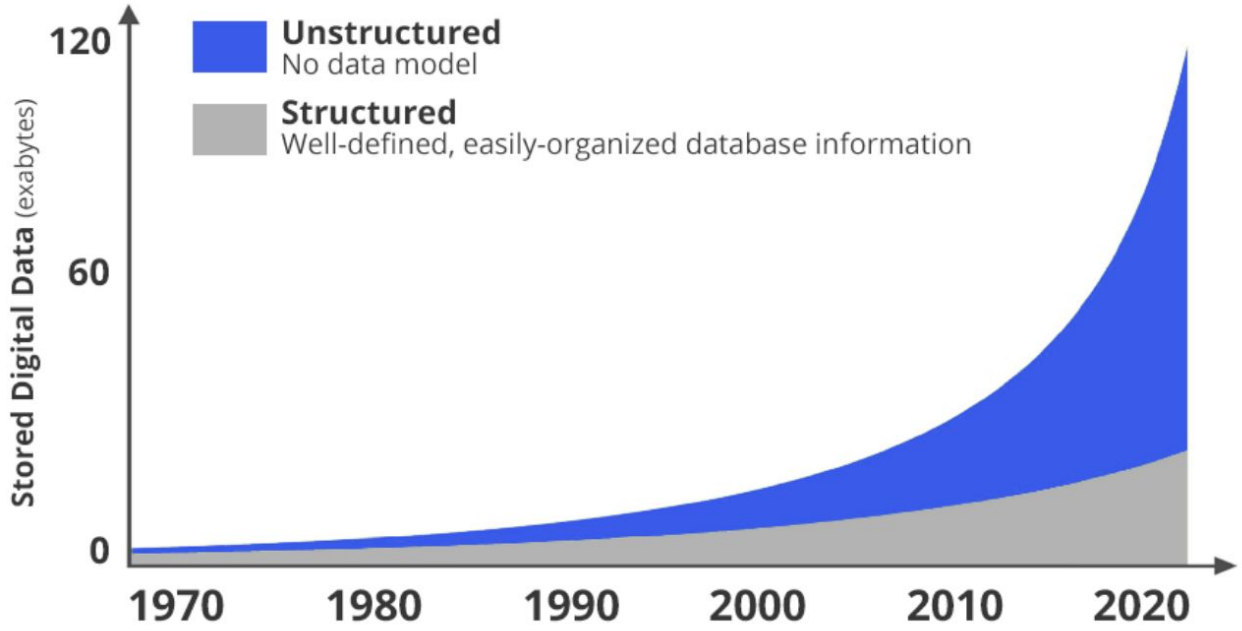
Motivations & Objectives

Motivations

- Exponential increase in the amount of data that we as a race are producing, particularly **unstructured data**.

- Researchers have employed the use of **lexicons** to systematically filter text throughout history.

- The popularisation of **word embeddings algorithms** have made the production of lexicons from large text corpora much more efficient



Objectives

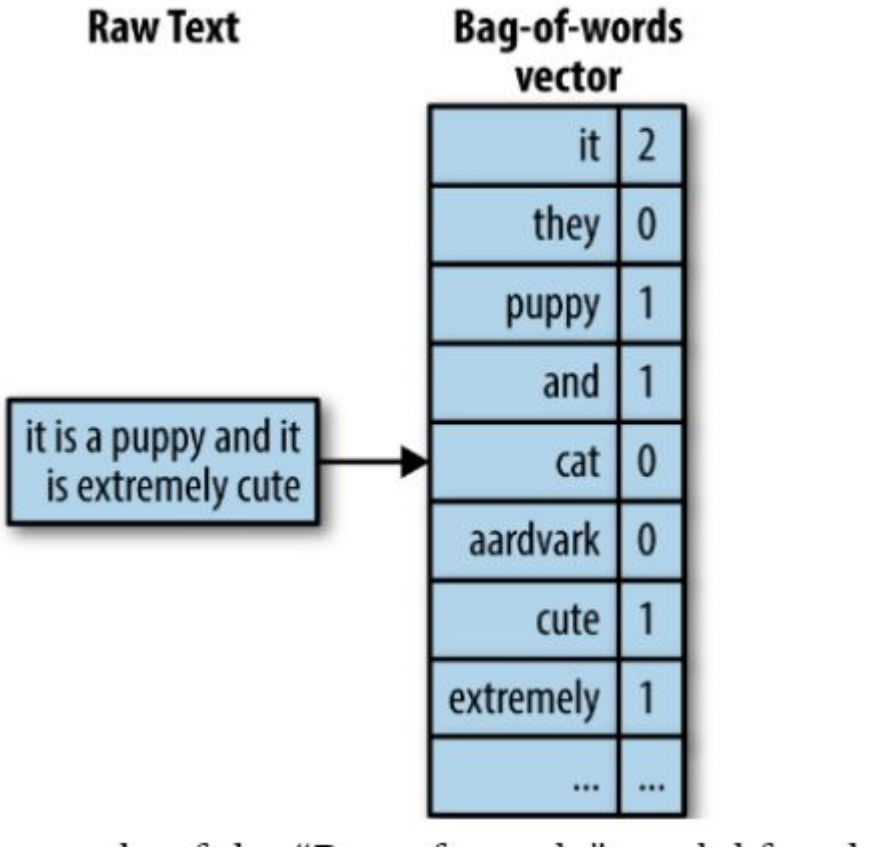
- Obtain a number of appropriate **large text datasets** of varying size and content.
- Create several different **word embedding models** using a combination of popular algorithms.
- **Compare and evaluate** pre-trained models.
- Develop an **ensemble recommendation method**
- Implement an **interactive web interface** for lexicon building, using the proposed lexicon recommendations.
- Conduct a **user evaluation** on the produced web application.

Related Work

Text Mining

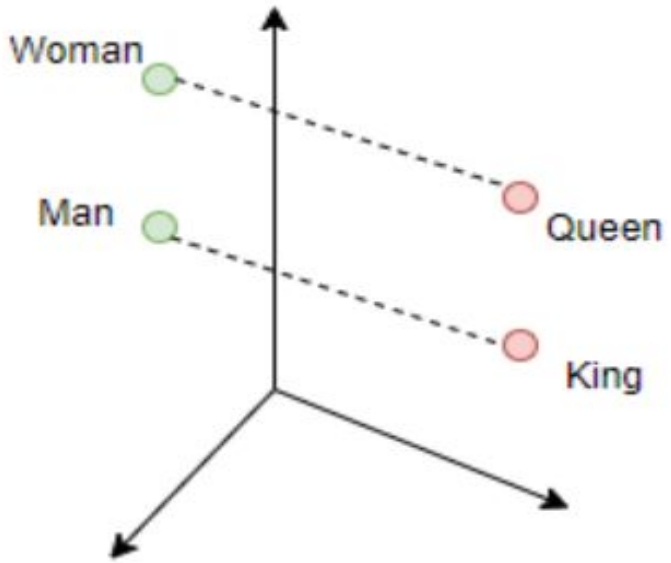
- The process of extracting useful information from unstructured textual data through the identification and exploration of interesting patterns.

- Four applications of Text Mining applicable:
  - Information Retrieval (IR)
  - Text Classification
  - Natural Language Processing (NLP)
  - Information Extraction (IE)

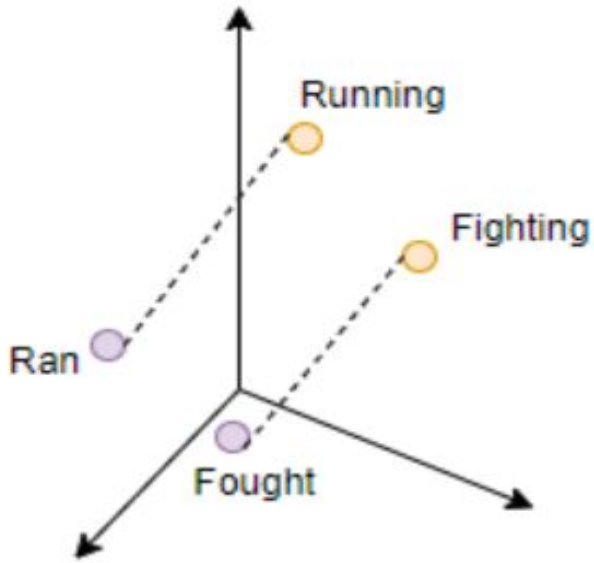


Word Embeddings

- Word Embeddings are a class of techniques where individual words are assigned and represented as **real-valued vectors in predefined vector space**.
- Each word “mapped” to one vector.
- Similar words appear close together in the vector space.
- Popular Algorithms:
  - Word2Vec
  - GloVe
  - FastText



Male-Female Relationship



Verb Tense

Lexicon Building

-

Data Considerations

Initial Datasets

Consisted of three initial datasets of various sizes and content.

Data Source	Google News	Reddit News	Wikipedia Dump
Size	3.5 GB	80 MB	10 MB
Pre-trained	Yes	No	Yes
Array Dimensions	300	100	32
Algorithm	Word2Vec	Word2Vec	Word2Vec

Future Datasets:

Create two self-trained models from competing political news outlets. Eg. Fox News + CNN

Evaluate Word2Vec Models against pre-trained GloVe (1.4GB) and FastText (650MB)

Outline of Initial Approach

Creation And Evaluation Word Embedding Models

A comparison of 3 different models produced from the seed word “fight”, outputting the top 7 most similar words.

Most Similar Google	Most Similar Reddit	Most Similar Wikipedia
fighting	battle	fights
fights	combat	enemies
battle	struggle	outnumbered
fought	drive	avenge
Fight	fighting	retaliation
bout	help	angered
battles	jihad	escalated

Future Work

Visualisation of Embedding Models

- T-SNE
- UMAP
- Visual Comparison of CNN and Fox
- Example Visualisation

Development of Ensemble Method

- Combining of Ranked List
- Rank Aggregation

Implementation of Web Application

- Flask
- Pickle
- Bokeh
- Visualisation

Evaluation of Web Application

- What type of study
- How many people
- What questions

Summary and Conclusions