Initial Plan

Algorithmic 'Idea' clustering via natural language processing

Final Year Project, CM3203, 40 credits

Author: Niall Curtis

Supervisor: Alun Preece

Table of Contents

Project Description

During the author's year in industry, the company he worked for (Simply Do Ideas) produced a web application for other companies in order to enable them to conduct challenge-based innovation, through crowdsourcing ideas from their employees. In any given challenge that a company offers, it may amass a very large number of idea contributions, which can be a problem for the innovation staff to sift through manually. While the web application offers methods for tagging and pinning ideas to categorise them, it still requires a not-insignificant amount of man hours to read all of them, many of which can be empty, incomplete or duplicates of others.

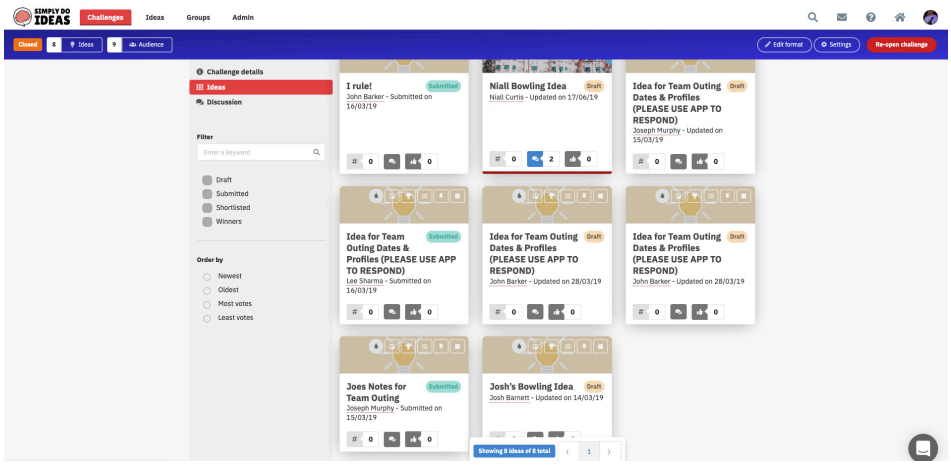

Figure 1 - Challenge with set of non-specific ideas

This project aims to address part of this problem by developing a post-ideation system for challenge owners that will streamline the process of reading submissions through automatic text analysis of ideas. The system in question would use appropriate algorithms and libraries to profile the corpus, with the main goal to cluster the ideas into groups of key challenge topics identified through tf-idf (Term Frequency – Inverse Document Frequency) profiling. It would also build

embedding models from the data, to conceptualise the virtual similarity/closeness of ideas in the data space.

The notion of clustering ideas (albeit in an argument/discussion environment) has been explored before, as seen in the report 'Visualised Clustering of Idea for Group Argumentation' (Luo 2010), where ideas from group discussion are classified intelligently then visualised. What makes the author's project state-of-the-art is furthering this idea to a more all-encompassing analytic system beyond a conceptual implementation to 'commercialise' it, as well as use more avant-garde exploratory analytics such as word embeddings. Additionally, focus is put on a visualisation implementation that is easily usable by somebody with weaker technical ability, making these advanced analytics accessible to the everyday user.

As well as building the underlying implementation to process the ideas, the project will focus on human computer interaction for the client; due to the aim being a faster and richer idea sifting experience, the speed and efficiency of user operation is as integral as the speed and intricacy of the underlying system. To this end, the idea analysis implementation will be paired with a well-justified user experience/interface, that will maximise the usability and usefulness of the generated metadata, with a suitable dynamic format to display the idea clusters and closeness.

In summary, this project will produce a system that can automatically pick out key topics of a dataset through analytic techniques, then visualise idea relationships.

Ethics

The project will be using anonymised ideas from real-world sources. These will ideas will be taken from consenting clients of Simply Do Ideas; with the knowledge they will be used for education/research. Personal information will be removed before submission/presentation to comply with legal & ethical requirements. As the author is a current employee of the company, he does not need to sign any NDA to access and manipulate this information, and the privacy policy of the organisation stipulates that anonymised data may be used for reporting in the future. After consultation with the author's supervisor with this information, they concluded that ethical information is not required as the dataset is already collected from existing sources and has necessary approval.

Project Aims and Objectives

The project can be split into two distinct sections, the system for analysing and clustering ideas, and the user interface for presenting and manipulating the produced data. Due to the nature of how interlinked they are, they will mostly need to be produced sequentially – Initial clustering and similarity exercises will need implementation before the data can be visualised. As such, the project can be

represented in both a “minimum viable” and “desirable” format, the latter of which comprises of stretch goals that increase the value of the project if time allows. The specific aims will be split into implementation and visualisation, then labelled if they are a ‘stretch goal’ for the desirable project.

Overall, the main project should produce a standalone system that takes in a JSON dataset of idea objects, performs the aforementioned analysis, and output another JSON file with the word frequency, TD-IDF matrices, cluster information and, ideally, word embeddings for the dataset. Additionally, a web-based UI will be able to reason the given dataset and visualise the information in a way that reduces the users’ burden of sifting through the entire corpus of ideas normally, enabling quick extraction of general trends and topics. For the purposes of the project, the UI will mostly reason with a static dataset and exist as a fungible black box implementation. In a commercial environment it would sit inside the existing web application and deal with a dynamic dataset.

Each task has been given a unique task ID (TX) which will be referred to in the timeline.

Key Supplementary Tasks

- **T1 - Data Clean-Up (MINIMUM):** System must be able to accept and normalise a large variety of ideas in different formats, to build a standardised corpus of ideas for further comparison, amalgamating the content for text processing.

Analyse

- **T2 – Word Frequency (MINIMUM):** Word and bigram frequency analysis; allowing simple visualisation of the most used terms. Both segmentation and tokenization will be used, along with a common and user-extended library of stop-words to remove analytically useless connectives, verbs and common terms that have no use for investigation
- **T3 – TD-IDF (MINIMUM):** Analytic techniques to pick out the most ‘popular’ terms in the set of ideas, which in turn allows discovery of the overall ‘key topics’. Acts as basis for clustering and enables broad stroke inspection of challenge outcomes.
- **T4 – Clustering (MINIMUM):** Act of separating the ideas into previously unknown, potentially non-discrete groups based on their individual TD-IDF matrices, traits, and the occurrent of trends within the ideas; the aim being to create subsets of ideas that belong to a particular topic within the idea corpus, which were previously found via inverse document frequency analysis. This is the end goal to solve the post ideation process of filtering ideas, as the administrators can immediately extract and dissect the key outcomes of the challenge.
- **T5 – Word Embeddings and Idea Closeness (DESIRABLE):** Representing individual words as vectors in a vector space, finding words that have the same representation and thus being able to define some degree of similarity, or closeness in the vector space, between different ideas. Using a pre-

existing populated embeddings model for word comparison, we can compare the features of ideas and be able to produce a traversable map of ideas.

Visualise

- **T6 – Key topics (MINIMUM):** Simple visualisation of the key topics acquired from TD-IDF and term frequency analysis data, to enable a minimum viable product
- **T7 – Clustering representation (MINIMUM):** A well-justified method to represent the clustering of ideas in the dataspace, that fully supports a fast and efficient idea sifting process for the user. Clearly define the extracted key topics and allow granular access to ideas clustered within these topics for deeper analysis.
- **T8 – Embedding graph (DESIRABLE):** Take advantage of the embedding vector-space and similarity data to create a linked map visualisation of the closeness of ideas. Can pick out trends between clusters, navigate the data space, view potential outliers and more.
- **T9 – Data manipulation (DESIRABLE):** Create set of tools that give the user the ability to manipulate the tools that analyse the dataset in real time to allow them to decide their own granularity of analysis – By empowering them to complete actions such as adjusting cluster size, with the visualisation robust enough to support these user activities.

Work Plan

The project will be split into 2-week sprint increments, that end with a supervisor meeting to discuss progress made over the fortnight towards the milestone of that period, which link to one of the task IDs defined in the last section. Throughout the plan are 3 intermittent stakeholder reviews, which will review the current state of the system at that time with the company, supervisor and any interested parties, which will allow iteration and improvement.

Week 1 – Supervisor Meeting

- Produce initial work plan with main objectives.
- Initial research into appropriate libraries and algorithms for system & exploratory analysis of test ideas using these.
- **Milestone: Proof of Concept and Initial Plan**

Week 2

- Preparation for system implementation, setting up Python environment with dependencies, data formats for input and output.
- Understand and document usage of chosen algorithms for data analysis.
- Implement method for cleaning and normalising corpus of ideas to allow text processing.
- **Milestone: Base system with methods to receive and normalise idea set (T1)**

Week 3 – Supervisor Meeting

- Term frequency analysis tool.

- Plan user interface implementation, gather requirements, wireframe components.
- **Milestone: System outputs corpus term frequency, set of necessary documents on wireframe (T2)**

Week 4

- Create TD-IDF matrix production tools for idea content, produce set of 'important' terms for data.
- Report on progress made in system using notes and adjust plan as necessary.
- **Milestone: System outputs TD-IDF matrices for dataset, final report created and partially filled, plan reverified (T3)**

Week 5 – Supervisor Meeting

- Implement initial UI from previously produced plans, with visualisation for TD-IDF and term frequency data from idea analysis system.
- Demonstration of base-level system to relevant stakeholders.
- **Desirable:** Plan and research word embeddings using word2vec.
- **Milestone: Front-end web application connected to analysis system, stakeholder demo completed (T6)**

Week 6

- Implementation for idea clustering, using the TD-IDF data for appropriate labels for clusters. Optimise and bug fix existing as necessary.
- **Milestone: System outputs a set of clusters for ideas based on key extracted topics (T4)**

Week 7 – Supervisor Meeting

- Increment existing user interface using visualisation library to demonstrate idea clusters, and allow broad stroke analysis with 'topics'.
- **Desirable:** Implement word embeddings model to calculate idea closeness **(T5)**
- **Milestone: Web application updated to reason with and visualise cluster data (T7)**

Week 8

- Testing system output, relevancy of results, appropriateness of interface and user experience – Necessary improvements implemented from this.
- **Desirable:** Visualise word embeddings with vector space graph **(T8)**
- **Milestone: Thorough documentation on the effectiveness of data analysis, efficiency and usefulness of visualisation**

Week 9 – Supervisor Meeting

- Begin producing rest of final report, collate notes produced earlier and plan as necessary.
- Demo of fully completed product to stakeholders.
- **Desirable:** Implement user controls for navigating data space and managing granularity **(T9)**
- **Milestone: Subsection of report produced with structure and plan in place to continue, stakeholder demo completed**

Week 10/11

- Complete main body of final report including full reporting on implementation, results, evaluation, conclusion.
- **Desirable:** Demonstration of desirable tasks **(T5, T8, T9)** to stakeholders
- **Milestone: Completed and reviewed final report document prepared for submission**

Week 12

- Editorial on final report, submission of project.
- **Milestone: Packaged submission with implementation and UI source, finished final report**

References

Luo, B. a. T. X. 2010. Visualized Clustering of Ideas for Group Argumentation. pp. 136-141. doi: 10.1109/WI-IAT.2010.280

School of Computer Science and Informatics

Coursework Submission Cover Sheet

Please use Adobe Reader to complete this form. Other applications may cause incompatibility issues.

Student Number	C1623580
Module Code	CM3203
Submission date	03/02/20
Hours spent on this exercise	20
Special Provision	

(Please place an x is the box above if you have provided appropriate evidence of need to the Disability & Dyslexia Service and have requested this adjustment).

Group Submission

For group submissions, *each member of the group must submit a copy of the coversheet.* Please include the student number of the group member tasked with submitting the assignment.

Student number of submitting group member	

By submitting this cover sheet you are confirming that the submission has been checked, and that the submitted files are final and complete.

Declaration

By submitting this cover sheet you are accepting the terms of the following declaration.

I hereby declare that the attached submission (or my contribution to it in the case of group submissions) is all my own work, that it has not previously been submitted for assessment and that I have not knowingly allowed it to be copied by another student. I understand that deceiving or attempting to deceive examiners by passing off the work of another writer, as one's own is plagiarism. I also understand that plagiarising another's work or knowingly allowing another student to plagiarise from my work is against the University regulations and that doing so will result in loss of marks and possible disciplinary proceedings.