Adv. IR Project Proposal
Hybrid Semantic-Lexical Framework for Document Clustering

Ben Ben Zvi 325526671 Noam Sasson 214439929

December 9, 2024

Problem Description

Document clustering relies heavily on inter-document similarity measures, which are typically based on unsupervised models like TF-IDF or BM25. While effective for capturing term frequency and inverse document frequency, these methods struggle to capture deeper semantic relationships. On the other hand, embeddings from large language models such as BERT and GPT offer rich contextual representations but lack sensitivity to corpus-specific distributions. This gap motivates the development of a hybrid similarity measure that integrates LLM embeddings with traditional unsupervised representations.

State of the Art

So far, cluster based retrieval used only classic statistical model representaions and similarity measures between documents. We will compare our results using different methods (score based, interpolation ext.) and compare our results to the models which got highest MAP score in their method.

Proposed Model

We propose a hybrid embedding framework that combines the semantic richness of LLM embeddings with the statistical grounding of unsupervised models, aiming to create a representation that balances global semantic patterns with corpus specific lexical nuances. This approach is expected to improve clustering performance, as it seeks to balance the trade off between semantic depth and lexical specificity, ensuring clusters are both contextually meaningful and sensitive to corpus specific variations, and offering a generalizable solution for clustering across diverse datasets.

Data

We will use the Reuters-21578 Text Categorization Collection to check our clustering performance, and tune hyperparameters accordingly. Then we will test our model on datasets such as TREC using one of the cluster based retrivals we saw in class.

Extra Details and Context

Formal Problem Definition

Given a set of documents $D = \{d_1, d_2, ..., d_n\}$ and 2 mappings $f1 : D \rightarrow \mathbb{R}^m$ and $f2 : D \rightarrow \mathbb{R}^n$ that represent the documents in 2 different spaces, we want to find a new mapping $f3(f1, f2) : D \rightarrow \mathbb{R}^k$ that combines the information from the 2 spaces in such way that will help to cluster the documents in the new space $f3$. We consider a mapping $f3$ to be a "good" mapping for clustering purpose if minimizing the intra-cluster distance in its space will also minimize the intra-cluster distance in the original spaces. Formally, let $C = \{c_1, c_2, ..., c_k\}$ be any given clustering of the documents (for simplicity assume hard clustering), the Within-Sum-of-Squares (WSS) of C in a space f is defined as $WSS(f, C) = \sum_{i=1}^{k} \sum_{d \in c_i} ||f(d) - \mu_{f,i}||^2$ where $\forall i \; \mu_{f,i} = \frac{1}{|c_i|} \sum_{d \in c_i} f(d)$ and are stricltly defined by the clustering C and the mapping f. the new mapping $f3$ should satisfy $\forall C \; WSS(f3, C) \propto \alpha WSS(f1, C) + \beta WSS(f2, C)$

Initial Solution

In our initial solution, we will show how simple concatenation of the two embeddings can statisfy the above condition. We will show that the new space $f3$ can be defined as $f3(d) = [\sqrt{\alpha} f1(d), \sqrt{\beta} f2(d)]$ and that the WSS of any clustering in this space is proportional (equal in this case) to the linear combination of the 2 WSSs in the original spaces.

$$\begin{gathered}
\alpha WSS(C, f1) + \beta WSS(C, f2) \\
= \alpha \sum_{c_i \in C} \sum_{d \in C_i} \|f_1(d) - \mu_i^{f1}\|_2^2 + \beta \sum_{c_i \in C} \sum_{d \in C_i} \|f_2(d) - \mu_i^{f2}\|_2^2 \\
= \sum_{c_i \in C} \sum_{d \in C_i} \sum_{j=0}^{m} (\sqrt{\alpha} f_1(d)_j - \sqrt{\alpha} \mu_{i,j}^{f1})^2 + \sum_{c_i \in C} \sum_{d \in C_i} \sum_{j=0}^{n} (\sqrt{\beta} f_2(d)_j - \sqrt{\beta} \mu_{i,j}^{f2})^2 = \dots \\
\text{defining } f^{new}(d) = \begin{pmatrix} \sqrt{\alpha} f_1(d) \\ \sqrt{\beta} f_2(d) \end{pmatrix} \text{ we get } \mu^{f_{new}} = \begin{pmatrix} \sqrt{\alpha} \mu_i^1 \\ \sqrt{\beta} \mu_i^2 \end{pmatrix} \text{ and:} \\
\dots = \sum_{c_i \in C} \sum_{d \in C_i} \sum_{j=0}^{m+n} (f^{new}(d)_j - \mu_{i,j}^{new})^2 \\
= \sum_{c_i \in C} \sum_{d \in C_i} \|f^{new}(d) - \mu_i^{f^{new}}\|_2^2 \\
= WSS(C, f^{new})
\end{gathered}$$

More Complex Problem Definition and Conection to Meta-Clustering

Given a set of documents $D = \{d_1, d_2, ..., d_n\}$, and n mappings $f_i : D \rightarrow \mathbb{R}^{m_i}$ that represent the documents in n different spaces, we want to find a mapping $f_{new}(f_1, f_2, ..., f_n) : D \rightarrow \mathbb{R}^k$ that combines the information from the n spaces in a way that will help to cluster the documents in the new space f_{new}. this problem can be seen as a generalization of the previous problem, where $n = 2$, and our solution for the simple problem can be generalized to this problem (show genral condition to good mapping for clustering purpose and show that the new space f_{new} can be defined as $f_{new}(d) = [\sqrt{\alpha_1} f_1(d), \sqrt{\alpha_2} f_2(d), ..., \sqrt{\alpha_n} f_n(d)]$).

Now we can redefine the problem to fit a meta-clustering problem, a previously explored problem in the literature. Given a set of documents $D = \{d_1, d_2, ..., d_n\}$, and n clustrings $C_1, \dots, C_n$ all derived from clustering in n different spaces $f_1, \dots, f_n$, we want to find a clustering C_{new} that best represents the information from the n clustrings. This is essentially the same setting and goal as the previous defenition, with the only difference being that the clustrings are given and not the mappings.

Related Work

Unsupervised models for clustering:
[1] Neepa Shah and Sunita Mahajan, PhD ”Document Clustering: A Detailed Review”
[2] Michael Steinbach, George Karypis and Vipin Kumar ”A Comparison of Document Clustering Techniques”

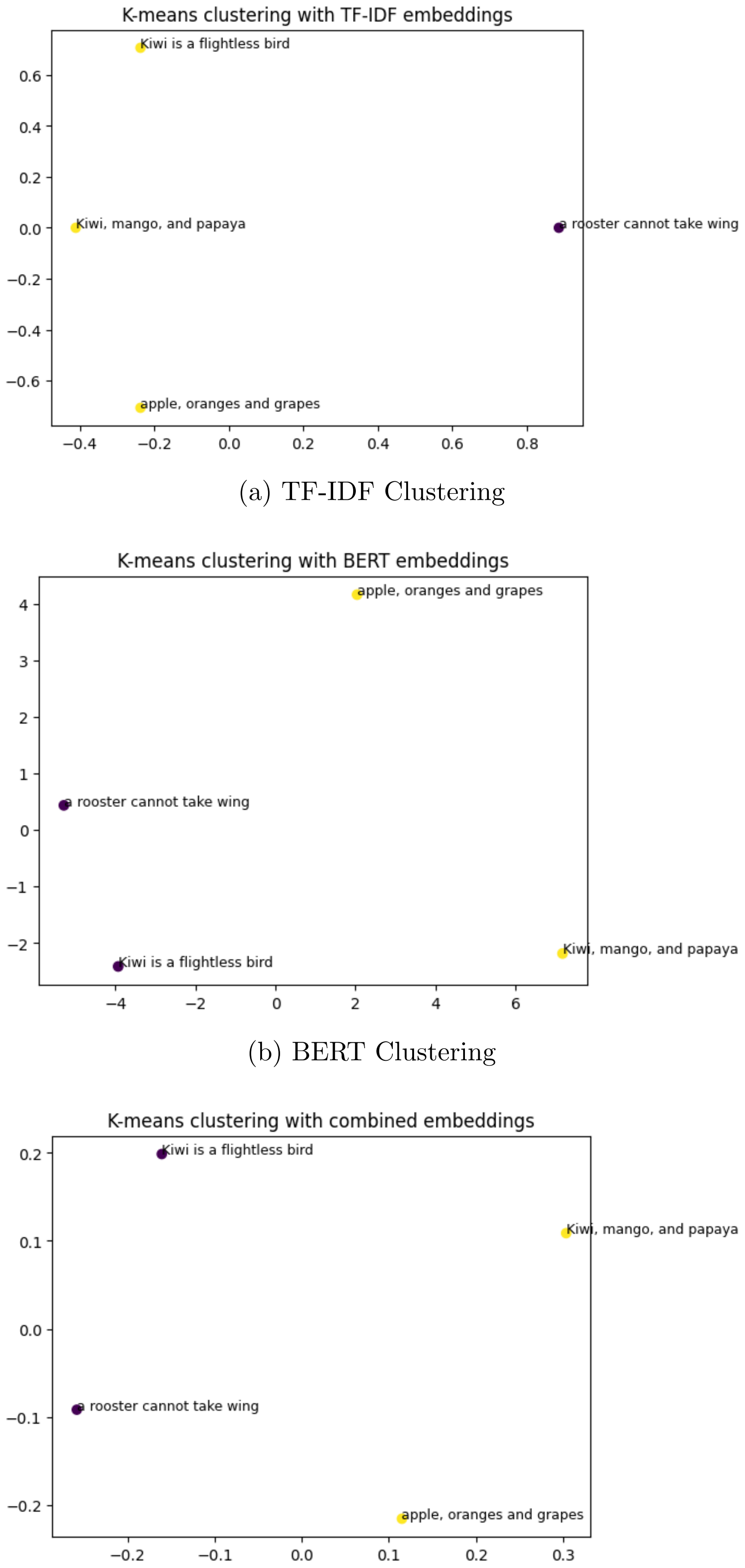
LLMs for document clustering:
[3] Alina Petukhova, João P. Matos-Carvalho and Nuno Fachada ”Text Clustering with Large Language Model Embeddings”

Meta-Clustering:
[4] Alexander Strehl and Joydeep Ghosh. ”Cluster ensembles - a knowledge reuse framework for combining multiple partitions”. Journal on Machine Learning Research

Simple Showcase

We will show a simple example for the simple problem definition, and our proposed solution, where $f1$ is a simple TF-IDF representation and $f2$ is Bert embeddings.
we chose $\sqrt{\alpha}$ and $\sqrt{\beta}$ to be $\sqrt{\frac{1}{\dim(f1)}}$ and $\sqrt{\frac{1}{\dim(f2)}}$ respectively, and used the combined embedding for clustering with K-means on 4 dummy documents:
d1 = ”Kiwi, mango, and papaya”
d2 = ”Kiwi is a flightless bird”
d3 = ”a rooster cannot take wing”
d4 = ”apple, oranges and grapes”

The results of the clustering can be seen in Figures 1a, 1b, and 1c.



(a) TF-IDF Clustering

(b) BERT Clustering

(c) Combined Clustering

Figure 1: Comparison of clustering methods: TF-IDF, BERT, and their combination.