Topic Modeling for Employee Objectives using Word Embeddings
Anum Afzal 01/02/2021
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion
Use case

CEO or HOD

Employee Objectives

Magic Box
Some common Topic Modeling approaches include LDA, LSA, PLSA
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion
Research Questions

RQ1: Could using embedding vectors lead to better results than Latent Dirichlet Allocation model?

RQ2: If the word embedding models are able to provide better results, then which type of embedding model is better suited?

RQ3: Could using a traditional algorithm such as LDA in tandem with the Embedding models provide better results?
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion
System Architecture of the Topic Modeling Framework that shows the interaction between the front-end and the back-end. 'Employee Analytic' tab reads meta-data from back-end before displaying options. It also fetches the requested results from back-end. 'Precompute Data' tab reads a new data file and generates results for all selected combination using the NLP engine and stores the result in back-end.
Agenda

- Overview
- Research Questions
- System Architecture

Datasets

- Methodology
- Results
- Discussion
- Demo
- Conclusion
<table>
<thead>
<tr>
<th>Column Name</th>
<th>Column Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>User ID</td>
<td>****0000</td>
</tr>
<tr>
<td>Global Key</td>
<td>HR</td>
</tr>
<tr>
<td>Functional Area</td>
<td>Human Resources</td>
</tr>
<tr>
<td>Location</td>
<td>Temecula</td>
</tr>
<tr>
<td>Objective Name</td>
<td>Operational Excellence</td>
</tr>
<tr>
<td>Objective Description</td>
<td>None</td>
</tr>
<tr>
<td>Objective Comment</td>
<td>I appreciate that Kathy has been continuing to work **** is a vital part of our support to our **** supports * information which is very much appreciated. I also ****, TOA to Sick Time and recently **** project among *. Thank You, Kathy! **** November 2019.</td>
</tr>
<tr>
<td>Objective Metric</td>
<td>Support **** and other activities focused around * and * for *. Support the ***, includes **, *** and other * along the way. Also expected is a regular ***** accurate * is recorded in ***** and is reflected in ** or selected tool. ***** initiatives ****. * and ***** the * of key insights for a further deep dive utilizing *. Continue to ** knowledge of ***** cluster. *** on a regular basis. Support efforts * from the ***** management and the * for ** Acknowledgement of same.</td>
</tr>
<tr>
<td>Form Template Name</td>
<td>2019 Performance Management</td>
</tr>
</tbody>
</table>

some information is redacted to anonymize the data for privacy reasons as this is a private dataset.
Employee Objective Dataset Analysis Summary

- Extract relevant columns
- Remove redundant objectives
- Objectives with english text
- Truncate at length 219
- Merge objectives belonging to same employee

~33,000 → 4078 → 3281 → 511
<table>
<thead>
<tr>
<th>Column Name</th>
<th>Column Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>12612628</td>
</tr>
<tr>
<td>Title</td>
<td>Engineering Systems Analyst</td>
</tr>
<tr>
<td>Full Description</td>
<td>Engineering Systems Analyst Dorking Surrey Salary ****K Our client is located in Dorking, Surrey and are looking for Engineering Systems Analyst our client provides specialist software development Keywords Mathematical Modelling, Risk Analysis, System Modelling, Optimisation, MISER, PIONEEER Engineering Systems Analyst Dorking Surrey Salary ****K</td>
</tr>
<tr>
<td>Location</td>
<td>Dorking</td>
</tr>
<tr>
<td>Contract Time</td>
<td>permanent</td>
</tr>
<tr>
<td>Contract Type</td>
<td>full_time</td>
</tr>
<tr>
<td>Company</td>
<td>Gregory Martin International</td>
</tr>
<tr>
<td>Category</td>
<td>Engineering Jobs</td>
</tr>
<tr>
<td>Salary</td>
<td>20000 - 30000/annum 20-30K</td>
</tr>
</tbody>
</table>
Job Description Dataset Analysis Summary

- Extract relevant columns
- Remove redundant descriptions
- Objectives with English text
- Truncate at length 709

>200,000 -> >200,000 -> >200,000 -> >200,000
Agenda

● Overview
● Research Questions
● System Architecture
● Datasets

● Methodology

● Results
● Discussion
● Demo
● Conclusion
Topic Modeling

Clustering

K = 1

K = 2

K = 5

Topic word retrieval

Clustering

K = 1

K = 2

K = 5
A Block diagram of the methodology used in this study. An objective document is passed through a Pre-processing step first, then one of five type of feature spaces is selected which is used of clustering. Next, one of the three topic word retrieval technique is selected to get the top topic words. Last step is post processing to remove redundant topics from clusters.
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology

- Results
- Discussion
- Demo
- Conclusion
Analytical Measures

1) **Coherence Score:**
   - Measures the degree of similarity between topics in a cluster.
   - Outputs a value between 0 and 1.

\[
\text{score}_{\text{UMass}}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}
\]

2) **Silhouette Score:**
   - Examines the compactness of the data point features within a cluster and how well the clusters are separated from each other.
   - Outputs a value between +1 and -1.

\[
s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}
\]
## Results
Experiments using Embedding Models on Job Description Dataset

<table>
<thead>
<tr>
<th>Embedding Model</th>
<th>Coherence Scores</th>
<th>Silhouette Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence BERT</td>
<td>0.5333</td>
<td>0.0638</td>
</tr>
<tr>
<td>BERT</td>
<td>0.5857</td>
<td>0.1321</td>
</tr>
<tr>
<td>XLNET</td>
<td>0.4938</td>
<td>0.1324</td>
</tr>
<tr>
<td>Sentence RoBERTa</td>
<td>0.5499</td>
<td>0.0427</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>0.5743</td>
<td>0.1169</td>
</tr>
<tr>
<td>Sentence DistilBERT</td>
<td>0.6040</td>
<td>0.0435</td>
</tr>
<tr>
<td>XLM</td>
<td><strong>0.6118</strong></td>
<td>0.0744</td>
</tr>
</tbody>
</table>

*Coherence scores and Silhouette Scores when using feature space from Embedding Models for clustering before post-processing step with 10 clusters, frequency-based topic word retrieval approach on Job Description dataset*
Coherence scores and Silhouette Scores when each feature space is used for clustering. Score are captured before post-processing step with 10 clusters, frequency-based topic word retrieval approach on Job Description dataset.

<table>
<thead>
<tr>
<th>Feature Spaces</th>
<th>Coherence Scores</th>
<th>Silhouette Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA model</td>
<td>0.4629</td>
<td>N/A</td>
</tr>
<tr>
<td>Embedding Model</td>
<td>0.5333</td>
<td>0.0638</td>
</tr>
<tr>
<td>Embedding Model + LDA</td>
<td>0.6001</td>
<td>0.0745</td>
</tr>
<tr>
<td>Embedding Model + Autoencoder</td>
<td><strong>0.6280</strong></td>
<td>0.1392</td>
</tr>
<tr>
<td>Embedding Model + LDA + Autoencoder</td>
<td>0.6083</td>
<td><strong>0.2456</strong></td>
</tr>
</tbody>
</table>

*Coherence scores and Silhouette Scores when each feature space is used for clustering. Score are captured before post-processing step with 10 clusters, frequency-based topic word retrieval approach on Job Description dataset.*
The questionnaire consists of 16 statements, divided into 8 evaluation aspects.
The participant read the statement and expressed their agreement/disagreement with the statement on a 1-5 scale: [Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree].
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion
RQ1: Could using embedding vectors lead to better results than Latent Dirichlet Allocation model?

Yes, using a Word Embedding model for Topic Model can lead to better results.

RQ2: If the word embedding models are able to provide better results, then which type of embedding model is better suited?

Sentence Transformers Models such as Sentence BERT, Sentence RoBERTa and Sentence DistilBERT provide the best results.

RQ3: Could using a traditional algorithm such as LDA in tandem with the Embedding models provide better results?

No, using Word Embedding in tandem with an LDA model provide almost the same result as using a Word Embedding model alone.
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion
Demo
Feature Space from LDA Model

\( k = 10 \) on Job Description dataset
Feature Spaces with Sentence BERT Embedding Model

K = 10, Job Description Dataset
Agenda

- Overview
- Research Questions
- System Architecture
- Datasets
- Methodology
- Results
- Discussion
- Demo
- Conclusion
Sentence Transformers such as Sentence BERT and Sentence RoBERTa provide an accurate feature space for clustering.

Clusters obtained using Embedding model is similar to the one obtained from feature concatenation and Autoencoder.

Silhouette score and Coherence score is not a good measure of evaluation for task such as Topic Modeling.

Quality dataset is essential for performing an unsupervised task such as Topic Modeling.
Thank you for your attention! Questions? Comments?

Anum Afzal
Bsc

Technische Universität München
Department of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

anum.afzal@in.tum.de
www.matthes.in.tum.de