Text Mining and Analysis

OBJECTIVE

SMART CV ANALYSIS

- In recent times, companies are integrating automation in areas which could increase the efficiency of the company by many-folds. In order to not be outdated, an on-line presence has become of significant nature.
- Building teams which can produce consistent results is what successful institutions strive for.
- Adapting to the changing nature, purpose of projects, the correct team can be assigned to the project by understanding the importance of the words in their resumes.

KEY OBSERVATIONS :

- Documents in which a particular word has a higher tf-idf score is more likely to get selected.
- Similarity of documents using cosine similarity and LDA document-topic matrix.

KEY LEARNINGS

TEXT MINING AND ANALYSIS

- Data used for analysis is typically structured.
    - Resumes is unstructured data. Therefore, structuring the data and viewing them as parameters is easier to analyze.
- Analysis Methods :

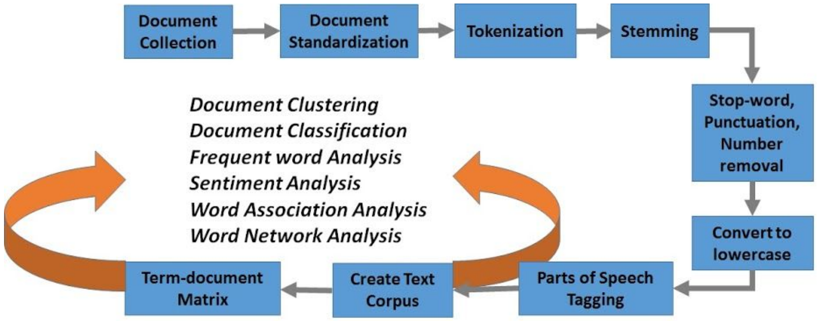
  (1) Bag of Words Model : The text is represented as the bag of its words , disregarding grammar and word order but maintaining multiplicity.

  (2) Document Term Matrix - description of frequency of terms the occur in the document.

  (3) TF-IDF : reflects how important the word is in the document, thus preventing the bias of term frequency.

  (4) Latent Dirichlet Allocation: Each document is a mixture of topics and each word is attributable to one of the topics.

  (5) Cosine Similarity : Determines the similarity between two documents based on the Euclidean distance of their respective vectors.
- Libraries used for implementation : NLTK,Scikit-Learn,pandas,numpy,pyplot.



Document Collection

Document Standardization

Tokenization

Stemming

Stop-word, Punctuation, Number removal

Convert to lowercase

Parts of Speech Tagging

Create Text Corpus

Term-document Matrix

Document Clustering
Document Classification
Frequent word Analysis
Sentiment Analysis
Word Association Analysis
Word Network Analysis

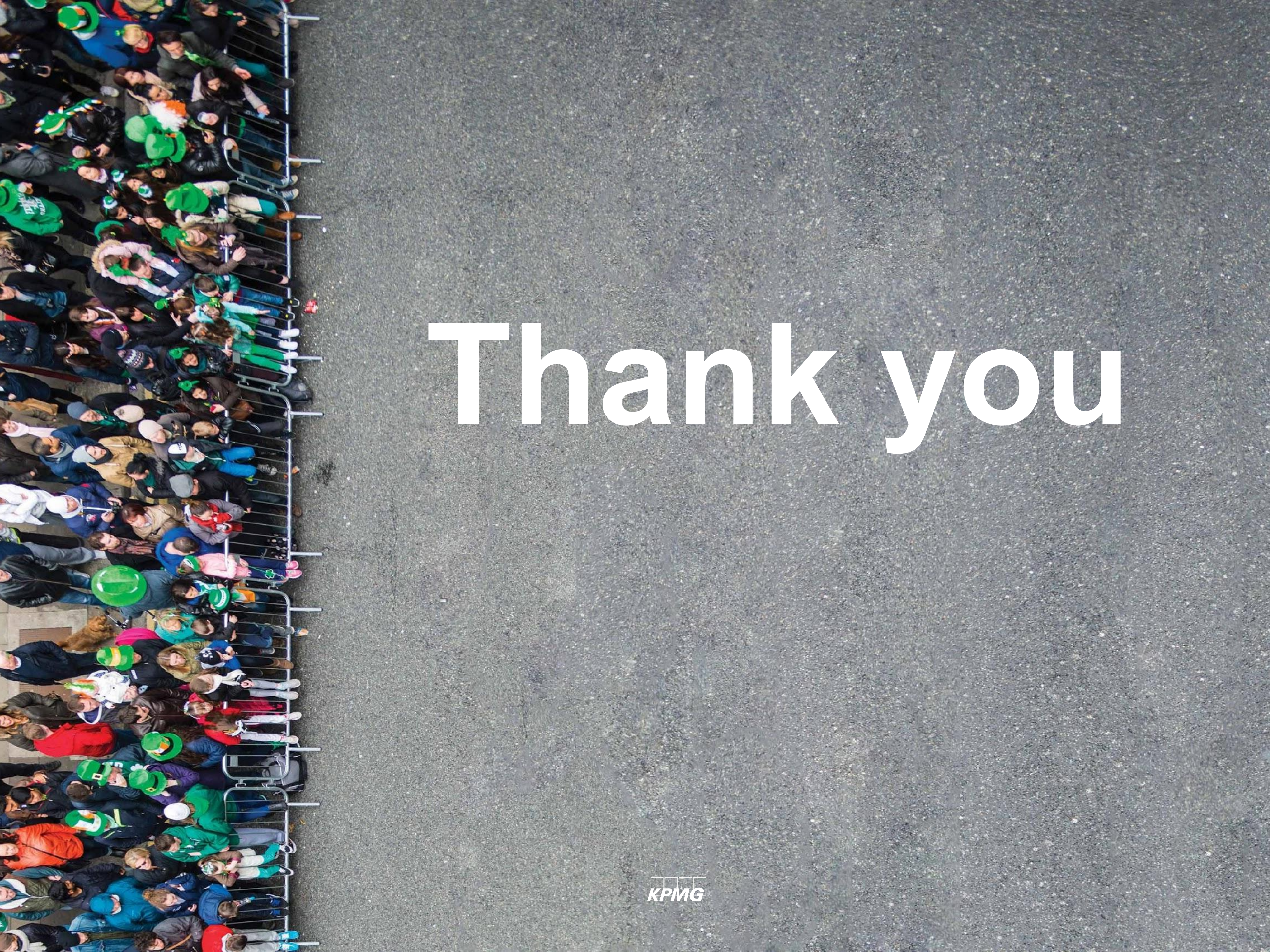
INFERENCE

TEXT ANALYSIS

- Higher the TF-IDF score , more important the word in the document.
- TF-IDF is a good metric to decide the importance of the word in the set of documents.
- LDA is an unsupervised method of learning. It results in abstract topics . Therefore, it is not a good method for topic modelling.

NEXT STEPS :

- Extending the model to larger dataset: To view a better implementation of the model , a larger dataset leads to more precise analysis.
- Matching the data with its respective fields. For example, matching the marks obtained 10th / 12th with the 10th / 12th fields under education.
- Correlation of the data : To achieve better evaluation of the data.



Thank you