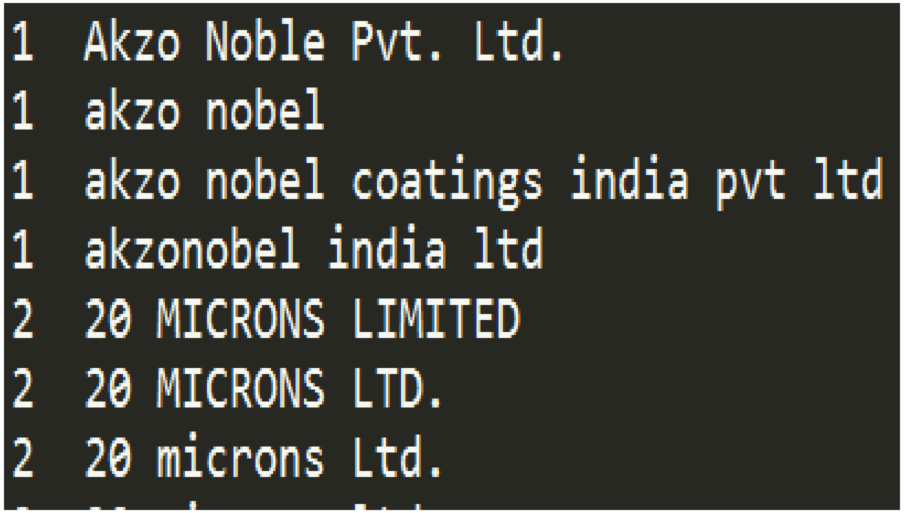
Using Entity Resolution to cluster names of organizations

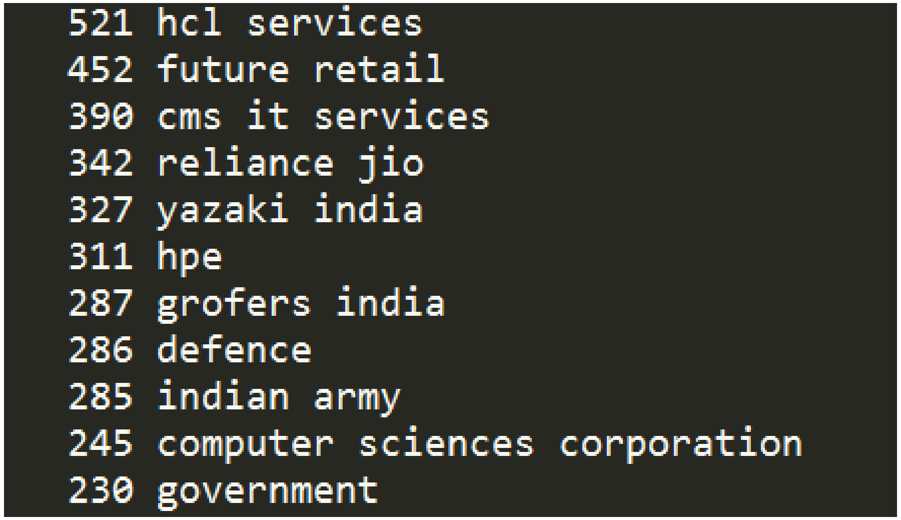
Data Description

- Company clusters (**Training set**)
- ***<cluster id, company name>***



```
1   Akzo Noble Pvt. Ltd.
1   akzo nobel
1   akzo nobel coatings india pvt ltd
1   akzonobel india ltd
2   20 MICRONS LIMITED
2   20 MICRONS LTD.
2   20 microns Ltd.
```

- Unmapped Samples (**Test Set**)
- ***<frequency, company name>***



```
521 hcl services
452 future retail
390 cms it services
342 reliance jio
327 yazaki india
311 hpe
287 grofers india
286 defence
285 indian army
245 computer sciences corporation
230 government
```

Problem Statement

- **Given:**
  - Company clusters sample (Training set)
    - ***<cluster id, company name>***
  - Unmapped Sample (Test Set)
    - ***<count of occurrence, company name>***

- **Target:**
  - Map the strings from the unmapped data file to a cluster id

- **Key points:**
  - The approach should be probabilistic.
  - In case the mapping is not possible, the variant file would need to be updated.

Approach

- **Preprocessing**
  - **Removed common terms** like “services”, special characters and white spaces
  - E.g. “citicorp services india” -> “citicorp”

- **Modeling**
  - Approach 1:
    - Clustering based on **common sequence of characters**
  - Approach 2:
    - Clustering based on **cosine similarity of feature vectors**

Approach : In detail

- **Approach 1**
  - String a = “Times Internet”; String b = “Times Internet Inc.”
  - Length of **longest common subsequence** is: 14

- **Approach 2**
  - String: “timesofindia”
  - Vector:

0	…	9	a	b	c	d	e	f	…	i	…	m	n	…
0	0	0	1	0	0	1	1	1	0	3	0	1	1	…

  - **Cosine similarity** between vectors [only the dimensions where either of them exists are considered]

The Code

- **Preprocessing (preprocess.py)**
  - https://github.com/sohomghosh/company_clustering/blob/master/preprocess.py

- **Modeling (model_code.R)**
  - https://github.com/sohomghosh/company_clustering/blob/master/model_code.R

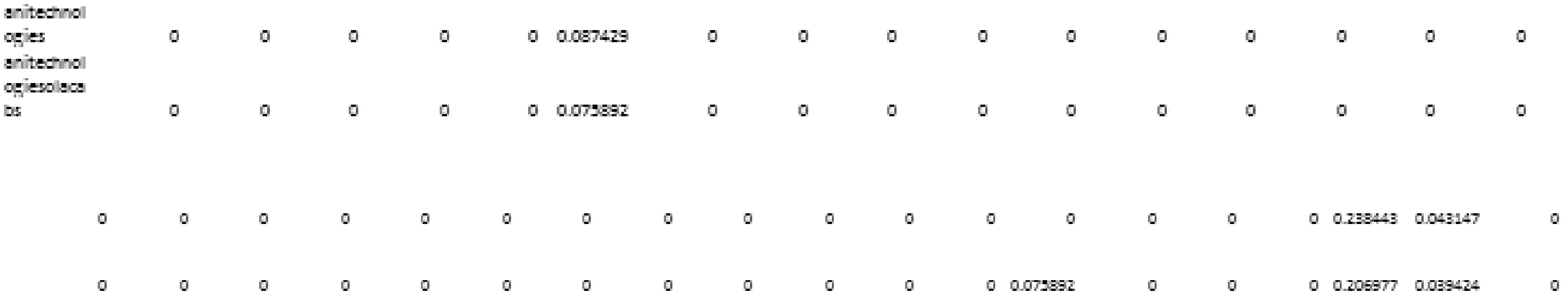
Results

- **Approach – 1**
  - updated_company_cluster_sample.csv
    - https://github.com/sohomghosh/company_clustering/blob/master/updated_company_cluster_sample.csv
  - cluster_distribution.csv
    - https://raw.githubusercontent.com/sohomghosh/company_clustering/master/cluster_distribution.csv

- **Approach – 2**
  - updated_company_cluster_sample_v2.csv
    - https://github.com/sohomghosh/company_clustering/blob/master/updated_company_cluster_sample_v2.csv
  - cluster_distribution.csv
    - https://raw.githubusercontent.com/sohomghosh/company_clustering/master/cluster_distribution_v2.csv

Discussion

- Companies like “**anitechnologies**”, “**anitechnologiesolacabs**” have similar cluster distributions



anitechnologies	0	0	0	0	0	0.087429	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.238443	0.043147	0
anitechnologiesolacabs	0	0	0	0	0	0.073892	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.073892	0	0	0	0.206977	0.039424	0

- **Challenges**
  - Setting a **optimal threshold**
  - **Scaling** the algorithm

Future Work : This Model

- **Removing white spaces and special characters** in the very first step makes it difficult for restoring back the original string.
- **Machine learning algorithms** for probabilistic classification may be used for assigning clusters.
- The output of the methods mentioned may be **ensembled** to produce better results.
- The **threshold** can be altered to **tune** the model further.
- The model needs to be trained on the **entire data** for increasing its efficiency.

Future Work : Other Models

- **Deep Learning based approach**
  - Training Deep Neural Network after extraction of features (Connected Entities, Relations, Entity Types, Entity Description) [Ref: 7]
- **Markov Logic based**
  - Combination of First Order Logic & Markov Networks [Ref: 8]
- **Graph based approaches using Map Reduce**
  - Blocking -> Linking -> Clustering [Ref: 6]
- And many others ….

How can I contribute?

- **Word Recommendation, Dynamic Analysis** of content of Ads being drafted at ads2book
- **Content summarizer** (120 words) for news articles, user reviews on jobs, interview questions etc.
- More **personalized news feeds, jobs** by integrating Social Media Data
- **Sentiment Analysis** of users’ reviews about jobs, news, movies etc.
- And much more to “*Simplify Life*” !!!

What I NEED? -> GUIDANCE & SUPPORT ☺

My Relevant Experience

- **Sentiment Analysis on Movie Reviews**
  - *[IJARCST, Vol 3, Issue 1, pp 41-46] (journal)*
- **Recommendation System based on Product Purchase Analysis**
  - *[ISSE, Springer London, ISSN:1614-5054, Vol 12, Issue 3, pp 177-192] (NASA journal)*
  - *[ICACNI, SIST Springer, ISBN: 978-81-322-2538-6, Vol 43, pp 581-591] (conference)*
- **Extraction & Analysis of Publication Data of Conferences**
  - *[IEEE International Conference on Advances in Computing & Communication Engineering-2015, pp 588-593]*
- **Analysis of Computer Science publications**
  - *[WIS & COLLNET 2015] (poster)*

References

1. http://www.cs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf
2. http://precog.iiitd.edu.in/Publications_files/Paridhi_Jain_Comprehensive_Report_Spring_2013.pdf
3. http://vldb.org/pvldb/vol5/p2018_lisegetoor_vldb2012.pdf
4. https://cran.r-project.org/web/packages/qualV/index.html
5. https://cran.r-project.org/web/packages/stringdist/stringdist.pdf
6. *H Kardes, D Konidena et. al,* ***Graph-based Approaches for Organization Entity Resolution in MapReduce***
7. *H Huang et. al,* ***Leveraging Deep Neural Networks and Knowledge Graphs for Entity Disambiguation***
8. P Singla, P Domingos, ***Entity Resolution with Markov Logic***

Questions?

Thank You