**********Publicly available datasets*******************************

1) DBpedia is a publically available ontology manually created based on the most commonly used infoboxes within Wikipedia.

Pairs of similar company variations has been extracted by running query on DBpedia SPARQL endpoint: http://dbpedia-live.openlinksw.com/sparql

*Sample query to check 100 name pairs :

```
select distinct ?s ?o
where {?s dbo:wikiPageRedirects ?o .
?s rdf:type dbo:Company .
?o rdf:type dbo:Company } limit 100
```

*All variations of a company in DBpedia are mapped to 1 single variation which can be used to group all the variations to entity clusters respectively.

The file DBpedia_clusters.tsv consists of the gold entity clusters.

2) ESCO is a publicly available dataset of clusters which can be downloaded from https://ec.europa.eu/esco/portal/home

From the website: ESCO (European Skills, Competences, Qualifications and Occupations) is the European multilingual classification of Skills, Competences, Qualifications and Occupations.

We downloaded ESCO's Skills/competences referred as "ESCO-Skill" in paper, and ESCO's Occupations referred as "ESCO-Designation" in paper.

The files esco_designation_ground_clusters.txt and esco_skills_ground_clusters.txt consists of gold clusters downloaded from website.

****Our contribution in preparing pairwise dataset for evaluating SM*******

We prepared esco_designation_pairs.csv, esco_skills_pairs.csv, DBpedia_company_pairs.tsv which consists of both similar and dissimilar pairs from open datasets using the method described in paper in Section 4.1. We plan to release this dataset which can be used to evaluate the pairwise similarity scores for variations by the research community.