Lab 2 results - Team 8

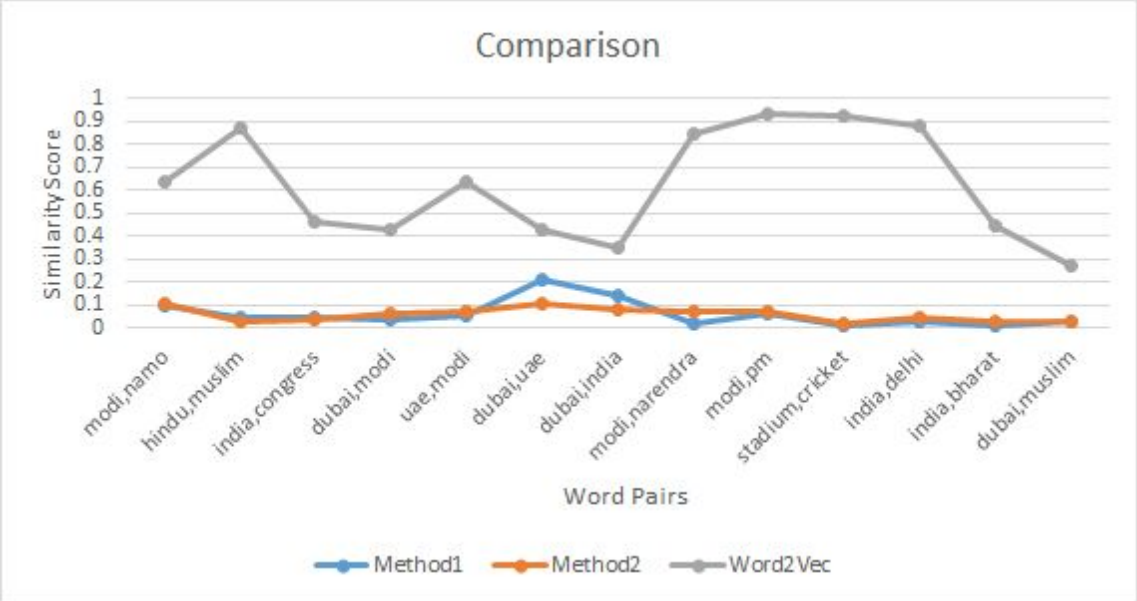
Observations-

- The formula for score was slightly modified:
  - **Method 1:** Z = (count of triplets where central words are w1 and w2 and the context for both are the same) / (count of all triplets where the central words are w1 and w2)
  - **Method 2:** Z = (count of triplets where central words are w1 and w2 and the context for both are the same) / (count of all triplets which had count of occurrences of the context of the words w1 and w2)
- This is a Naïve approach which is heavily dependent on the training data, i.e. the appearance of the words in the context in the training data.
- Some observations in tweets are:
  - There are a lot of spelling errors. eg.: "Thanks to Mr. Modi all my ungles and aunthies in Abu Dubhai will finally go to their own temble. Thiz iz zimbly zuperb"
  - Other languages are transliterated to English. eg.: "kullam khulla becho desh ko"
  - Extracting a tweet is hard as it spans several lines arbitrarily in certain cases.
  - Short forms are used in certain scenarios which is not present abundantly enough in the corpus to train for such occurrences.
- The word2vec with a window size of 3 was trained to serve as a baseline for comparison of the aforementioned methods.

Tabular comparison-

Word1	Word2	Method1	Method2	Word2Vec
modi	namo	0.09865623	0.10227473	0.638935852
hindu	muslim	0.04784689	0.02673797	0.874555651
india	congress	0.04106776	0.03870968	0.459788167
dubai	modi	0.03448877	0.05998182	0.424940284
uae	modi	0.05004284	0.07312797	0.641518941
dubai	uae	0.21018277	0.10128971	0.430534462
dubai	india	0.14393939	0.08072838	0.352826119
modi	narendra	0.01594836	0.07266436	0.844844673
modi	pm	0.05871915	0.06735751	0.934485338
stadium	cricket	0.00995025	0.0141844	0.928930266
india	delhi	0.02364865	0.04745763	0.883130499
india	bharat	0.01335559	0.02898551	0.44447244
dubai	muslim	0.02556539	0.02425373	0.269948413

Graphical plot -



- Its clearly evident from the graph that both the mentioned methods follow word2vec in most cases.
- The choice of the normalization defines the trend the method follows.
- The normalization with respect to the occurrences of the context performs better in most scenarios.

We also obtained a 0 similarity score for unrelated pairs of words such as:

```
Enter w1:bangkok
Enter w2:bahubali
Similarity(Z_score) = (0, 0)
Enter w1:relationship
Enter w2:cricket
Similarity(Z_score) = (0, 0)
Enter w1:
```