An application of Spectral Clustering on Named Entity Recognition

Ali Josue Limon, Michele Cerú

New York University

ajl649@nyu.edu; mc3784@nyu.edu

December 7, 2016

Project description

- Goal:
- Purpose:

Dataset

- bleble
- bleble

Model parameters

- Size of the word embedding space: d_E
- technique used for the word embedding training: cbow or skip-gram
- Number of words embedding considered around the target word: n_w
- Kernell used for the spectral clustering: cosine similarity
  - Cosine similarity
  - Knearest neighbours
  - $k(x, y) = exp(-\gamma||x - y||_1)$

Evaluation Criteria

Adjusted Rand index: accuracy of the predicted labels given the ground truth.

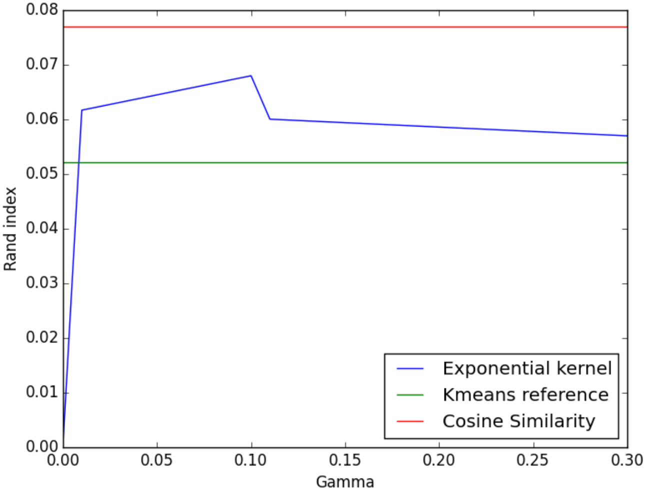
$$S_{11} = \{\text{pairs in the same cluster in C and C'}\}$$
$$S_{00} = \{\text{pairs in different cluster both in C and C'}\}$$
$$S_{10} = \{\text{pairs in the same cluste C but not in C'}\}$$
$$S_{01} = \{\text{pairs in the same cluste C' but not in C}\}$$

Rand index:

$$R = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} = \frac{n_{11} + n_{00}}{\binom{n}{2}}$$

Adjusted Rand index: Center the random clustering to 0

result

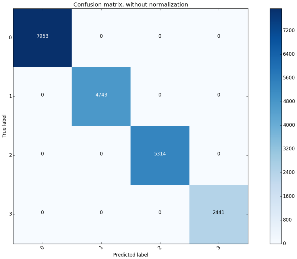


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Rand index
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Gamma
Exponential kernel
Kmeans reference
Cosine Similarity

Data samples distribution ground truth

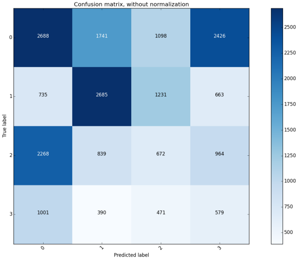
Cluster Label	Name Entity	Number of sample
0	Organization	7953
1	Location	4743
2	Person	5314
3	Miscellaneous	2441

Confusion matrix



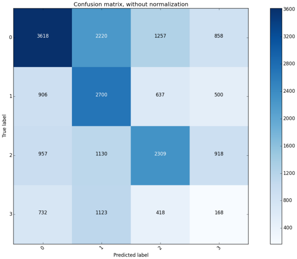
Confusion matrix, without normalization

	0	1	2	3
0	7953	0	0	0
1	0	4743	0	0
2	0	0	5314	0
3	0	0	0	2441



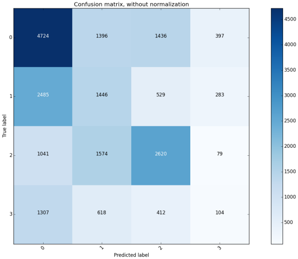
Confusion matrix, without normalization

	0	1	2	3
0	2688	1741	1098	2426
1	735	2685	1231	663
2	2268	839	672	964
3	1001	390	471	579



Confusion matrix, without normalization

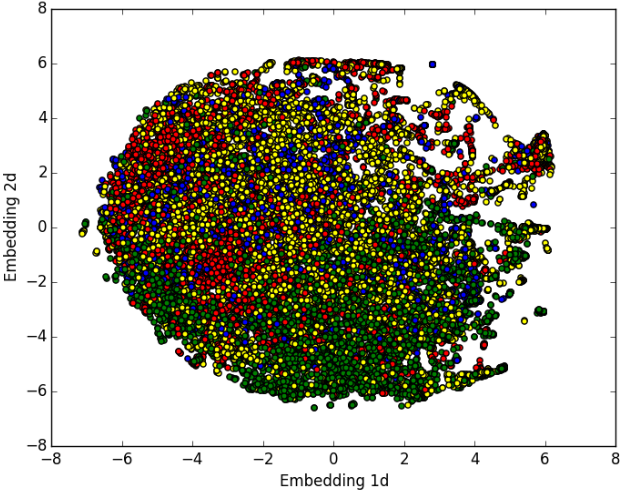
	0	1	2	3
0	3618	2220	1257	858
1	906	2700	637	500
2	957	1130	2309	918
3	732	1123	418	168



Confusion matrix, without normalization

	0	1	2	3
0	4724	1396	1436	397
1	2485	1446	529	283
2	1041	1574	2620	79
3	1307	618	412	104

Data visualisation in 2d



Conclusion

- bleble
- bleble