PCA for Short-text Keyword Detection

Methods

Trying dimensionality reduction algorithms for keyword extraction.

Each word is represented as a d-dimensional representation using word2vec. All the stop words were removed, and each document is represented as an n x d matrix. Once we get the components, for example, principal components after PCA, for each component, we check the value closest in cosine space, and that word is counted as a keyword.

The procedure is detailed below:

1) Remove stop words.
2) Applying PCA and take all the components (since the number of words will be small, important words would be repeating in the components. To avoid that, we are taking all the components)
3) For each component,
    a. Find the closest word in the original n x d matrix (cosine distance used)
    b. Add the word to the list of keywords
4) If a word is already added in the keyword list, do not add it to the list. Go to the next component.
5) If the number of words in keyword list is equal to the number of principal components, break.

Example Sentence:

Deoxyribonucleic acid (DNA) is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses.

Manually Labeled Ground Truth:

Deoxyribonucleic, acid, DNA, molecule, genetic, instructions, growth, development, functioning, reproduction, living.

Results are shown below:

Distance Metric	Number of Keywords Predicted	Number of Keywords in Ground Truth	Number of Matches
Euclidean	4	10	2
Cosine (Min)	9	10	6
Cosine (Max)	8	10	7
Correlation (Min)	9	10	6
Correlation (Max)	9	10	7

Chebyshev (Max)	1	10	1
Chebyshev (Min)	6	10	4
Braycurtis (Min)	8	10	6
Braycurtis (Max)	5	10	4
Canberra (Max)	1	10	1
Canberra (Min)	7	10	5

Another Example:

The process by which green plants and some other organisms use sunlight to synthesize foods from carbon dioxide and water. Photosynthesis in plants generally involves the green pigment chlorophyll and generates oxygen as a byproduct.

Results are shown below:

Distance Metric	Number of Keyword Predicted	Number of Keywords in Ground Truth	Number of Matches
Cosine (Min)	13	14	9
Cosine (Max)	10	14	6
Correlation (Max)	10	14	6
Correlation (Min)	13	14	9

Hits and Misses:

Ground Truth	Predicted (Position)	Hit/Miss
Process	Process (13)	Hit
Plants	plants (3)	Hit
Sunlight	sunlight (5)	Hit
Synthesize	synthesize (4)	Hit
Foods	foods (6)	Hit
Carbon		**Miss**
Dioxide	dioxide (10)	Hit
Water	water (1)	Hit
Photosynthesis	Photosynthesis (11)	Hit
Green		**Miss**
Pigment	pigment (7)	Hit
Chlorophyll	chlorophyll (2)	Hit
Oxygen		**Miss**
Byproduct		**Miss**

Other predicted words include involves (8), generates (9), and generally (12).