**Assignment Description**

**Semantic Similarity/Relatedness Measure between words**

The study of semantic similarity between words has been a part of natural language processing and information retrieval for many years. Semantic similarity is a generic issue in a variety of applications in the areas of computational linguistics and artificial intelligence, both in the academic community and industry. Examples include word sense disambiguation, detection and correction of word spelling errors (malapropisms), text segmentation, image retrieval, multimodal documents retrieval, and automatic hypertext linking. Similarity between two words is often represented by similarity between concepts associated with the two words.

To decide whether two words are semantically similar, it is important to know the semantic relations that hold between the words. For example, the words horse and cow can be considered semantically similar because both horses and cows are useful animals in agriculture. Similarly, a horse and a car can be considered semantically similar because cars, and historically horses, are used for transportation.

Accurate extraction of semantically related words requires a precise definition of the closeness between a pair of words, in terms of either the pair wised similarity or distance. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, F-score, Euclidean distance, Mahalanobis distance, and the Jaccard coefficient and Pearson Correlation Coefficient.

### METHOD BASED ON PEARSON CORRELATION COEFFICIENT

The Pearson’s Correlation coefficient between *i*th and *j*th words is calculated by the following formula:

The steps involved in our proposed method are:

**Step 1:** Pre-processing of input data

**Step 2:** Semantically related terms extraction

Each of the steps is described in detail in the following sections.

Step 1: Pre-processing of input data

Each document is supposed to transform into some text format, one of the major problems in text mining is that a document can contain a very large number of words. If each of these words is represented as a vector coordinate, the number of dimensions would be too high for the text mining algorithm. Hence it is crucial to apply preprocessing methods that greatly reduce the number of dimensions (words) to be given to text mining algorithms. Our method also apply several pre processing methods to the original documents, namely stemming, removal of stop words and proper nouns, and elimination /drop of very low and very high frequent words.

**Stemming**

Stemming of keywords involves extracting the root words, dropping such things as plurals and gerunds(ing). Some examples would include: running/run, apples/apple, and educational/ education. Most search engines perform stemming by default when providing results so that a search on something like "green apples" will not exclude sites that only use the singular form, "green apple".

**Removal of Stopwords**

Stopword also known as close-class words (i.e. insignificant words like ‘can’, ‘in’, ‘this’, ‘from’, ‘then’, ‘or’, ‘the’, ‘by’) are some of the stopwords they occur very frequently in a document. Since they are so common in many documents, they carry very little information about the contents of a document in which they appear. Most Search Engines do not consider extremely common words in order to save disk space or to speed up search results.

This English stopword list was built by Gerard Salton and Chris Buckley for the experimental SMART information retrieval system at Cornell University. This stopword list is generally considered to be on the larger side and so when it is used, some implementations edit it so that it is better suited for a given domain and audience while others use this stopword list as it stands. This wordlist is 720 words in length.

**Elimination of very High and very Low Frequency Terms/Words**

Luhn proposed that the frequency of the word occurrence in an article furnishes a useful measurement of word significance. Luhn used Zipf's Law as a null hypothesis to specify two cut-offs, an upper and a lower, thus excluding non-significant words. The words exceeding the upper cut-off were considered to be common and those below the lower cut-off rare, and therefore not contributing significantly to the content of the article. The advantages of the process are non-significant words are removed so that they will not interfere during retrieval, also the size of the total text can be reduced.

**Illustration**

Let us take a football news article.

|  |
| --- |
| **Original**: A night that promised much drama ended in disappointment for Scotland as a 3-1 defeat in Spain left them third in Group I and denied them a chance of a place in the Euro 2012 play-offs. Two first half goals from Manchester City's David Silva and a David Villa strike were enough for Spain who top the group with a 100 percent record from eight matches. Scotland's consolation came from a David Goodwillie penalty on 64 minutes after Victor Valdes had tripped Steven Naismith. With Scotland's nearest rivals the Czech Republic winning 4-1 in Lithuania in a game played at the same time, Scotland needed an unlikely victory to keep their qualification hopes alive. The home team were determined to show why they are reigning European and World Champions and settled into their passing game straight from kick-off, indeed it was two minutes before a Scottish player touched the ball. As early as five minutes Spain were ahead. Valencia's Jordi Alba, on his debut, found space on the left wing to centre for Silva who finished crisply first time inside the near post. And by the 15th minute when the Czechs were two goals up in Lithuania, one a penalty awarded by a Spanish referee, it was clear Scotland were going to get little from their trip to Alicante. Indeed so complete was Spain's dominance, it wasn't until the 19th minute that they conceded a foul. On 34 minutes Darren Fletcher came close for Scotland with a left footed shot and when four minutes later James Morrison shot high from a wide left position Scottish hopes were momentarily revived. The Spanish however continued to create and Allan McGregor in Scotland's goal had to make a great outstretched save on 41 minutes from Santi Cazorla. Three minutes later he could do nothing when Silva cut in from the right wing, exchanged passes with Pedro Rodriguez, and shot into the left hand corner to send the Scots in needing a second half miracle. On 53 minutes Cazorla, Silva and Xavi Hernandez linked up delightfully on the right to find Villa in space in the Scottish area and the Barcelona striker easily slotted away his 50th goal in international football. Scotland continued to toil and created more in the second half with substitute Goodwillie in particular providing some much needed spark in attack. A fine performance from McGregor at the other end meant Spain were not to add to their tally, but 3,000 Scottish fans in the stadium and thousands more who travelled to the Spanish coast were left to wonder when their team would qualify again for their first major tournament since the 1998 World Cup in France. |

|  |
| --- |
| **Stemmed**: [A, night, that, promise, much, drama, end, in, disappointment, for, Scotland, as, a, 3, -, 1, defeat , in, Spain, left, them, third, in, Group, I, and, deny, them, a, chance, of, a, place, in, the, Euro, 2012, playoff, Two, first, half, goal, from, Manchester, City, David, Silva, and, a, David, Villa, strike, were, enough, for, Spain, who, top, the, group, with, a, 100, percent, record, from, eight, match, Scotland, console, come, from, a, David, Goodwillie, penalty, on, 64, minute, after, Victor, Valde, had, trip, Steven, Naismith, With, Scotland, nearest, rival, the, Czech, Republic, win, 4, -, 1, in, Lithuania, in, a, game, play, at, the, same, time, Scotland, need, an, unlike, victory, to, keep, their, qualify, hope, alive, The, home, team, were, determine, to, show, why, they, are, reign, European, and, World, Champion, and, settle, into, their, pass, game, straight, from, kickoff, indeed, it, was, two, minute, before, a, Scottish, player, touch, the, ball, As early, as, five, minute, Spain, were, ahead, Valencia, Jord, Alba, on, his, debut, find, space, on, the, left wing, to, centre, for, Silva, who, finish, crisply, first, time, inside, the, near, post, And, by, the, 15, th, minute, when, the, Czech, were, two goal, up, in, Lithuania, one, a, penalty, award, by, a, Spanish, referee, it, was, clear, Scotland, were, go, to, get, little, from, their, trip, to, Alicante, Indeed, so, complete, was, Spain, dominant, it, was, not, until, the, 19, th, minute, that, they concede, a, foul, On, 34, minute, Darren, Fletcher, came, close, for, Scotland, with, a, left, foot, shot, and, when, four, minute, later, James, Morrison, shot, high, from, a, wide, left, position, Scottish, hope, were, moment, revive, The, Spanish, however, continue, to, create, and, Allan, McGregor, in, Scotland, goal, had, to, make, a, great, outstretch, save, on, 41, minute, from, Santi, Cazorla, Three, minute, later, he, could, do, nothing, when, Silva, cut, in, from, the, right, wing, exchange, pass, with, Pedro, Rodriguez, and, shot, into, the, left, hand, corner, to, send, the, Scots, in, need, a, second, half, miracle, On, 53, minute, Cazorla, Silva, and, Xavi, Hernandez, link, up, delightful, on, the, right, to, find, Villa, in, space, in, the, Scottish, area, and , the, Barcelona, striker, easy, slot, away, his, 50, th, goal, in, international, football, Scotland , continue, to, toil, and, create, more, in, the, second, half, with, substitute, Goodwillie, in , particular, provide, some, much, need, spark, in, attack, A, fine, performance, from, McGregor, at, the, other, end, mean, Spain, were, not, to, add, to, their, tally, but, 3 , 000, Scottish, fan, in, the, stadium, and, thousand, more, who, travel, to, the, Spanish, coast, were, left, to, wonder, when, their, team, would, qualify, again, for, their, first, major, tournament, since, the, 1998, World, Cup, in, France] |
| **Without Stopwords**:[night, promise, drama, disappointment, Scotland, defeat, Spain, Group, deny, chance, place, Euro, 2012, playoff, goal, Manchester, City, David, Silva, David, Villa, strike, Spain, top, group, 100, percent, record, match, Scotland, console, David, Goodwillie, penalty, 64, minute, Victor, Valde, trip, Steven, Naismith, Scotland, rival, Czech, Republic, win, 4, -, 1, Lithuania, game, play, time, Scotland, need, victory, qualify, hope, alive, home, team, determine, show, reign, European, World, Champion, settle, pass, game, straight, kickoff, minute, Scottish, player, touch, ball, minute, Spain, ahead, Valencia, Jord, Alba, debut, space, wing, centre, Silva, finish, crisply, time, inside, post, 15, minute, Czech, goal, Lithuania, penalty, award, Spanish, referee, clear, Scotland, trip, Alicante, complete, Spain, dominant, 19, minute, concede, foul, 34, minute, Darren, Fletcher, Scotland, foot, shot, minute, later, James, Morrison, shot, high, wide, position, Scottish, hope, moment, revive, Spanish, continue, Allan, McGregor, Scotland, goal, great, outstretch, save, 41, minute, Santi, Cazorla, minute, Silva, cut, wing, exchange, pass, Pedro, Rodriguez, shot, hand, corner, Scot, miracle, 53, minute, Cazorla, Silva, Xavi, Hernandez, link, delightful, Villa, space, Scottish, area, Barcelona, striker, slot, 50, goal, international, football, Scotland , toil, substitute, Goodwillie, spark, attack, fine, performance, McGregor, end, mean, Spain, tally, 3 , 000, Scottish, fan, stadium, travel, Spanish, coast, wonder, team, qualify, major, tournament, 1998, World, Cup, France] |

|  |
| --- |
| **Without Proper Noun, numeral and special characters:** [night, promise, drama, disappointment, defeat, Group, deny, chance, place, playoff, goal, City, strike, top, group, percent, record, match, console, penalty, minute, trip, rival, win, game, play, time, need, victory, qualify, hope, alive, home, team, determine, show, reign, World, Champion, settle, pass, game, straight, kickoff, minute, player, touch, ball, minute, ahead, debut, space, wing, centre, finish, crisply, time, inside, post, minute, goal, penalty, award, referee, clear, trip, complete, dominant, minute, concede, foul, minute, foot, shot, minute, shot, high, wide, position, hope, moment, revive, continue, goal, great, outstretch, save, minute, minute, cut, wing, exchange, pass, shot, hand, corner, miracle, minute, link, delightful, space, area, striker, slot, goal, international, football, toil, substitute, spark, attack, fine, performance, end, mean, tally, fan, stadium, travel, coast, wonder, team, qualify, major, tournament, World, Cup] |

After above process each document is represented as bag of terms and then we find **distinct terms / Unique word** list in all documents.

**Step 2: Semantically related terms extraction**

The steps for extracting semantically related terms using Pearson correlation coefficient distance measure are given below.

**Step 2.1:** Generate the frequency matrix ***F*** for m documents and ***n*** distinct terms.

*F* =

Where ***fij*** is the no of occurrences of term ***tj*** in document ***di***

**Step 2.2:** Calculate the Pearson correlation coefficient using the formula

…(4.1)

where *i*≠*j* = 1,2,3…*n*.

**Step 2.3:** Combine the semantically related terms on the basis of Maximum Correlation Coefficient between pair of words.

The maximum correlation coefficient can be calculated by using the pseudo code in Algorithm 1 (described below).

**Algorithm 1:** find\_max\_correlation()

|  |  |
| --- | --- |
| **Input** | **:**  **R:** Correlation Matrix |
| **Output** | **:** Maximumcorrelation in **R** |
| **1.** | Let ***n*** is number of unique commands |
| **2.** | **Do for** *i* = 1 to *n*-1 |
| **3.** | *maxcoff* = *r11* |
| **4.** | **Do for** *j* *= i*+1 **to** |
| **5.** | **If**( *maxcoff*<*rij*) **Then** |
| **6.** | *maxcoff* = *rij* |
| **7.** | **Endif** |
| **8.** | **EndDo** |
| **9.** | **EndDo** |

**Illustration**

Given below is list of 5 pre-processed documents on football news. The terms present in each document is given row-wise.

|  |  |
| --- | --- |
| **Documnet1:** | [Coach, play, referee, goal, game, goal, mistake, linesman, referee, line, win, coach, mistake, unhappy] |
| **Document2:** | [score, goal, match, meet, World, Cup, quarter, final, win, goal, tournament, score, win, people ] |
| **Document3:** | [quarter, final, World, cup, ball, midfield, beat, hit, ball, score, goal, score, referee, offside] |
| **Document4:** | [World, Cup, quarter, final, score, goal, score, beat, team, extra, time, referee, score, quarter, game, win, final] |
| **Document5:** | [World, Cup, game, lose, win, tournament, match, goal, team, lose, semi, final, World, Cup, tournament, win, game, semi final] |

**Step 2.1:** The list of ***distinct terms*** in all five documents is,

|  |
| --- |
| [Coach, play, referee, game, mistake, ball, win, World, Cup, quarter, final, tournament, team, extra, time, win, line, meet, score, lose, semi, people, match, unhappy linesman] |

After removing those terms which have less than 2 occurrences in all documents the list is

|  |
| --- |
| [Referee, World, cup, tournament, quarter, final, score, win, game, goal] |

The frequency matrix F generated for the above sample data is

F =

**Step 2.2:** Calculate the correlation coefficient using the above formula in equation (4.1) for all pair of distinct terms. Let us assume that we want to calculate the correlation coefficient between the pair of terms say t1 and t2 then the formula will be

= f1,1× f2,1 + f1,2× f2,2 + f1,3× f2,3 + f1,4× f2,4 + f1,5× f2,5

= 2×1 + 0×0 + 1×0 + 1×1 + 0×1

= 2+0+0+1+0= 3 ...(4.2)

= f1,1 + f1,2 + f1,3 + f1,4 + f1,5

= 2+0+1+1+0= 4 ...(4.3)

= f2,1 + f2,2 + f2,3 + f2,4 + f2,5

= 1+0+0+1+1= 3 ...(4.4)

From equations (4.3), and (4.4) -

= =2.40 ...(4.5)

= (f1,1)2 + (f1,2)2 +(f1,3)2 +(f1,4)2 +(f1,5)2

= (1)2 + (2)2 +(1)2 +(1)2 +(0)2

= 1 + 4 + 1 + 1 + 0 = 6 …(4.6)

From equation (4.3) –

= = = 3.20 …(4.7)

= (f2,1)2 + (f2,2)2 +(f2,3)2 +(f2,4)2 +(f2,5)2

= (1)2 + (0)2 +(1)2 +(1)2+(0)2

= 1+0 +1+1+0 = 1 ... (4.8)

From equation (4.4) –

= = = 0.20 ...(4.9)

Now from equations (4.2), (4.5), (4.4), (4.6), (4.7), (4.8), and (4.9) -

= = = =0.33

Hence Corr(1,2) = 0.33

Similarly Correlation Coefficient for all pair of distinct terms can be calculated and represented by matrix R of size n×n.

R =

**Step 2.3:** Combine the pair of terms using Pearson Maximum Correlation Coefficient (Table 4.1)

**Table 4.1:** Pearson Maximum Correlation Coefficient

|  |  |
| --- | --- |
| **Maximum**  **Correlation Coefficient** | **Related terms** |
| **1.00** | tournament, quarter |
| **0.95** | final, game |
| **0.79** | score, final, game |
| **0.79** | win, tournament, quarter |
| **0.56** | cup, win, tournament, quarter |
| **0.41** | World, cup, win, tournament, quarter |
| **0.22** | Referee, goal |

**Algorithm 2:** To create Word Pattern strings from training dataset

|  |  |
| --- | --- |
| **Input** | : Training data |
| **Output** | :Correlation Coefficient Matrix R, Pattern String set *S* |
| **1.** | Let *U* is set of unique words in training data |
| **2.** | Let *n* is number of unique words |
| **3.** | Create frequency matrix *F* |
| **4.** | **Do** for all |
| **5.** | **Do** for each |
| **6.** | **Do** for each |
| **7.** | *= Cal\_*Correlation(*F*, *i*, *j*) |
| **8.** | *maxcorr* = find\_max\_correlation() |
| **9.** | **If** (*stringcount*< *threshold*) **Then** |
| **10.** | add |
| **11.** | increament *stringcount* by 1 |
| **12.** | **Else** |
| **13.** | increment *k* by 1 for next string |
| **14.** | **EndIf** |
| **15.** | **EndDo** |
| **16.** | **EndDo** |
| **17.** | **EndDo** |

**Assignment Summary**

1. Preprocess the given Text Documents (i.e. do stemming, remove stop words).
2. Find list of unique words in all documents.
3. Find frequency of each unique word in all text documents.
4. Remove very high frequent and very low frequent words from the list of unique words.
5. Use remaining list of unique words and generate frequency matrix.
6. Calculate Pearson Correlation Coefficient between each pair of unique words.
7. Find the pair with maximum correlation coefficient.
8. Add this pair of words as a single word at the end of frequency matrix.
9. Eliminate the columns of these two words from the frequency matrix.
10. Repeat Step 6 through 9 till matrix exhaust.

**Please Note: 1. All OOPS students of Group- B are required to implement step 1- 6 during the next OOPS lab (Wednesday) October 8, 2014.**

**2. All OOPS students of Group- A are required to implement step 1- 7 during the next OOPS lab (Thursday) October 9, 2014.**

**3. Implementation of remaining steps 8- 10 should be completed by October 15-16, 2014.**