Text Analysis of IBM Cloud-Computing Articles

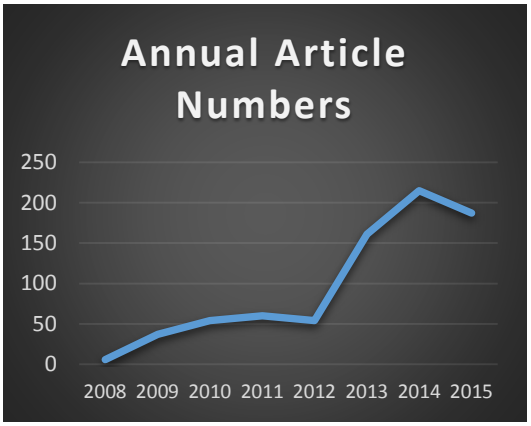
Nan Wang

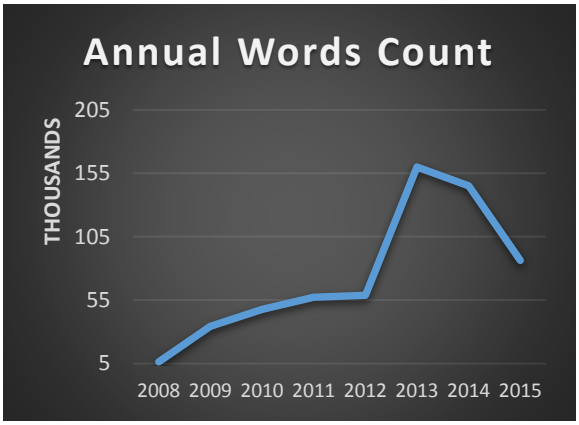
A. Dataset Description:

Use Scrapy, an open source and collaborative framework for extracting the data, to crawl all "Cloud-Computing-related" news articles, posted on http://www-03.ibm.com/press/us/en/pressreleases/recent.wss from the year 2008 through 2015.

Overall statistical summary of dataset:

	2008	2009	2010	2011	2012	2013	2014	2015
Article#	6	37	54	60	54	161	215	187
Word#	6394	34416	47700	57199	58857	159747	145055	86262





s

B. Project Goal:

To find out annual "hot word" related to IBM cloud computing from the year 2008 through 2015.

C. Methodology Summary:

1) Use a simple bag-of-words approach
2) Use a bag-of-words with stemming and stop word removal approach
3) Use a bag-of-words with stop word and customized word removal approach
4) Use POS and focus on NNP approach
5) Customized Python scripts to filter out common frequency words among years and find out unique frequency words for a particular year.

D. Analysis:

1) Simple bag-of-words approach:

The output is unsurprisingly terrible. With so many meaningless stop words, it is quite hard to find out the differences of frequent words between years, as is shown by two printouts below.

Frequent words for 2015:

```
In [6]: content_lower = [w.lower() for w in text]
        content = [w for w in content_lower if w.isalpha()]
        frdist_for_each_word = FreqDist(content)
        rank_each_word = sorted(frdist_for_each_word.items(), key = lambda x:x[1],reverse = True)
        rank_each_word

Out[6]: [('and', 2808),
         ('the', 2717),
         ('to', 1946),
         ('of', 1686),
         ('ibm', 1188),
         ('a', 1004),
         ('in', 986),
         ('for', 810),
         ('cloud', 665),
         ('with', 653),
         ('data', 493),
         ('is', 492),
         ('on', 485),
         ('that', 460),
         ('more', 413),
         ('as', 381),
         ('will', 361),
         ('new', 356),
         ('services', 334),
         ('from', 323)
```

Frequent words for 2014:

```
Out[12]: [('and', 4622),
          ('the', 4351),
          ('to', 3648),
          ('of', 2696),
          ('ibm', 2260),
          ('a', 1851),
          ('in', 1745),
          ('cloud', 1605),
          ('for', 1390),
          ('with', 1216),
          ('on', 935),
          ('is', 898),
          ('that', 848),
          ('more', 831),
          ('data', 778),
          ('new', 761),
          ('as', 694),
          ('will', 650),
          ('business', 540),
          ('it', 525)
```

2) bag-of-words approach with stemming and stop words removal：

The output improved compared to the first attempt. However, since the stemming package is so aggressive that it distorts meaning of many words, as shown in the printout below.

```
Which year you are gonna explore?
2015

Out[4]: [(u'ibm', 1190),
         (u'cloud', 685),
         (u'data', 493),
         (u'servic', 460),
         (u'new', 356),
         (u'develop', 315),
         (u'manag', 279),
         (u'busi', 272),
         (u'help', 264),
         (u'technolog', 259),
         (u'provid', 242),
         (u'watson', 240),
         (u'use', 233),
         (u'compani', 227),
         (u'inform', 216),
         (u'custom', 213),
         (u'solut', 209),
         (u'client', 207),
         (u'includ', 205),
         (u'platform', 194),
         (u'global', 182),
         (u'said', 179),
         (u'secur', 178),
         (u'applic', 176),
         (u'analyt', 164),
         (u'research', 159),
         (u'offer', 152),
         (u'enabl', 149),
         (u'system', 149),
         (u'innov', 138),
         (u'integr', 136),
         (u'health', 134),
         (u'enterpris', 133),
         (u'market', 131),
         (u'open', 130),
         (u'mobil', 129),
         (u'build', 129),
         (u'infrastructur', 129),
         (u'softwar', 128),
         (u'deliv', 126),
         (u'product', 125),
         (u'comput', 125),
         (u'capabl', 124),
         (u'storag', 123),
         (u'industri', 123),
         (u'center', 123),
         (u'visit', 118),
         (u'across', 116),
         (u'also', 115),
         (u'percent', 114),
         (u'oper', 113),
         (u'billion', 111),
         (u'app', 110),
```

3) bag-of-words approach with stop words and customized word removal:

Since all the articles are known as related to the topic of IBM Cloud Computing, it is highly possible that the three words, “IBM” “Cloud” “Computing” would appear frequently in articles posted in all of the eight years, which makes them meaningless if ranked as frequent words.

4) Use POS and focus on NNP approach:

The output is not as good as expected, for the reason that proper noun overall turns out to be not as frequently appear as other words in articles posted by IBM. The count of their appearance is too small to prove the significance of those words, as shown in the printout below.

```
Here are NNP words in articles written in 2015:

                    word  count
0        (zealand, NNP)      5
1         (watson, NNP)      5
2              (x, NNP)      3
3           (xaas, NNP)      2
4        (october, NNP)      2
5       (december, NNP)      2
6       (november, NNP)      1
7          (medal, NNP)      1
8       (ebeweber, NNP)      1
9        (griffin, NNP)      1
10            (xu, NNP)      1
11     (indians, NNPS)      1
12    (september, NNP)      1
13       (ryerson, NNP)      1
14           (yes, NNP)      1
15     (franklin, NNP)      1
16        (zurich, NNP)      1
17  (mobilefirst, NNP)      1
18      (kraemer, NNP)      1
19           (xiv, NNP)      1
20         (dixon, NNP)      1
```

5) **Filter out common frequency words among years and find out unique top frequency words for a particular year.**

After doing stop words and customized words removal (based on analysis above), the TFIDF idea is employed in this step: if a frequent word in one particular year also appears in other years, the importance weight of that frequent word decreases in that particular year. (Here, words ranked in top fifty in terms of frequency are regarded as frequent words.)

By comparing frequent words of every two out of eight (2008 to 2015) years, it is found that every year shares a large proportion of frequent words with other years, especially the year 2010 (with all of its 50 frequent words appearing also in other years).

	2008	2009	2010	2011	2012	2013	2014	2015
2008		27	26	25	24	25	25	22
2009	27		43	28	29	31	30	30
2010	26	43		31	32	33	33	33
2011	25	28	31		37	36	31	29
2012	24	29	32	37		40	35	36
2013	25	31	32	36	40		36	35
2014	25	30	33	31	35	36		37
2015	22	30	33	29	36	35	37	
All	22	48	50	44	46	48	45	45

(Numbers in the table above mean the counts of common frequent words in every two years.)

Year	Unique Frequent Words
2008	[well, president, partner, nc, vcl, virtual, state, sametime, unyte, idataplex, north, users, students, foundations, appliance, project, without, carolina]
2009	[web, processes]
2010	NONE
2011	[sales, commerce, application, government, key, public]
2012	[marketing, online, trademarks, puresystems]
2013	[provides, big]
2014	[portfolio, innovation, build, softlayer, us]
2015	[bluemix, weather, health, digital, billion]

E. Conclusion:

The output is not bad, since the intuition of words differ from year to year. It's reasonable to make a guess that IBM cloud computing business has shifted its target from government institution (as words "president" "state" "students" "Carolina" appeared a lot in early years) to commercial market("marketing", "trademarks") in the middle of time period concerned and, in the end, towards technological innovation field("innovation", "digital","softlayer").

In addition, it is interesting to note that the word of "big" stands out in the year 2013, which may indicates the popularity of "Big Data". Plus, It is worth mentioning that Bluemix, which is a hot IBM business nowadays, was first mentioned in the year 2015 with another word "health", both of are known as IBM core business.

F. Limitations:

Since stemming is not applied in this project, many words with plural form are considered as different from those carrying exact some meaning but in singular forms. It is necessary to transform all plurals into singulars (or vice versa), in the future analysis, to more accurately locate frequent words.

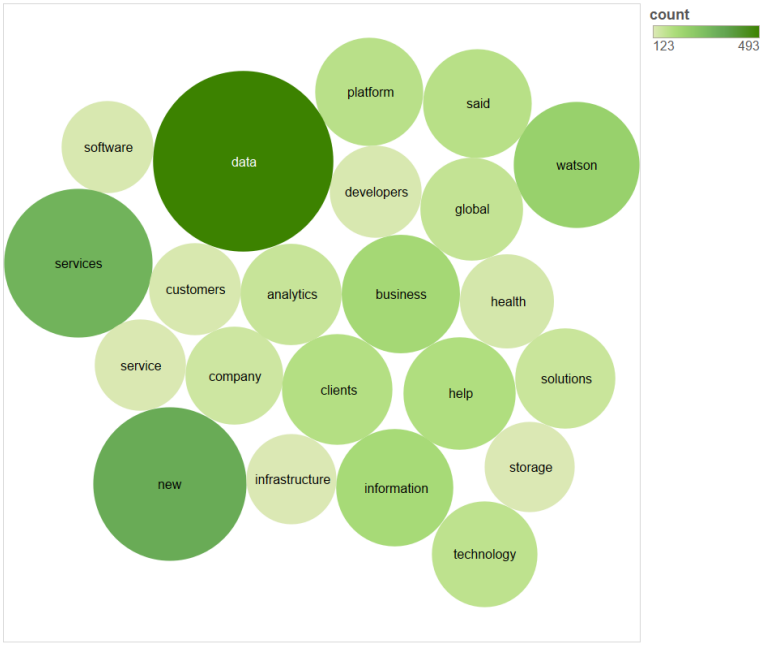
G. Future work:

To find words carrying the paradigmatic or syntagmatic relation with unique frequent words for each year. Sentiment analysis approach could also be applied, but with the knowledge that IBM highly advocates Cloud Computing, the result of sentiment analysis is doubtful to be informal.

H. Appendix:

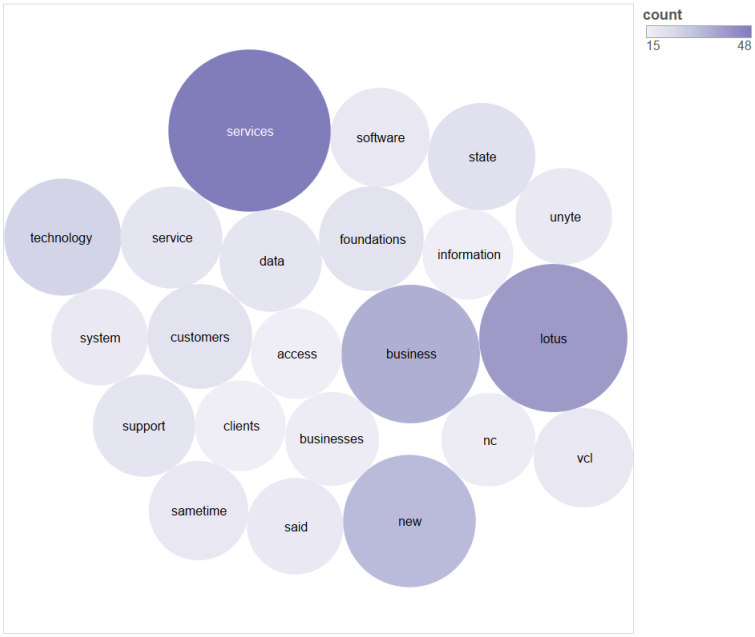
Visualization of frequent word distribution in the form of word cloud:

word cloud for 2015 IBM articles



Word. Color shows sum of count. Size shows sum of count. The marks are labeled by word. The view is filtered on sum of count, which ranges from 120 to 493.

word cloud for 2008 IBM articles



Word. Color shows sum of count. Size shows sum of count. The marks are labeled by word. The view is filtered on sum of count, which ranges from 15 to 48.