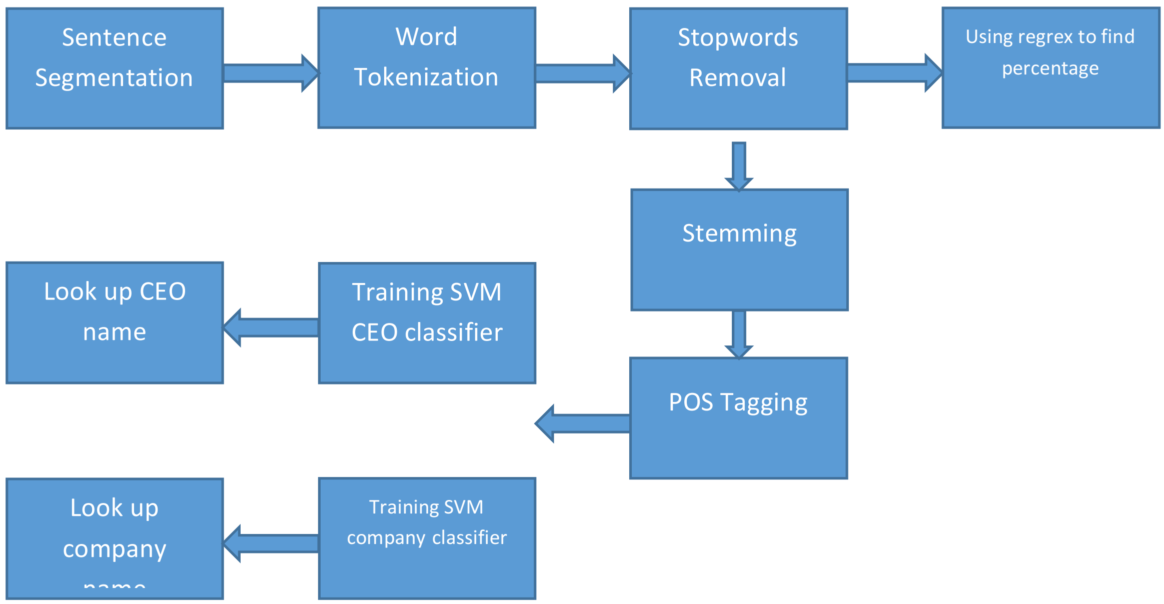
Zhaocheng yu netid zyr531

Target: extracting CEO names , company names and percentages.

Pipeline :



1.Preprocess

1. Sentence Segmentation:

   In this step, I just add all raw materials up into one file and then use Python Nltk package to finish the sentence segmentation. All relevant codes are in the sentence_seg.py. Output of this step is output.txt

```
REUTERS/China DailyHSBC China PMI fell to 50.5 in December, from 50.8 the previous
month.
A reading below 50 indicates contraction.
But this was right in line with expectations for 50.5, which was the flash PMI print.
"The moderation of December's final HSBC China Manufacturing PMI was mainly due to
slower output growth," said Hongbin Qu, HSBC chief economist China, in a press
release.
"However, the final PMI sustained the fifth above-50 reading in a row thanks to a
steady increase of new orders.
The recovering momentum since August 2013 is continuing into 2014, in our view.
With inflation still benign, we expect the current monetary and fiscal policy to
remain in place to support growth."
We got China's official PMI data on Tuesday, which showed manufacturing fall to 51,
from 51.4 in November.
The slowdown in credit growth is said to have weighed on Chinese manufacturing.
Here's the trajectory of Chinese manufacturing: HSBC/Markit Economics FREE
AppDownload

Netflix users and investors may have a hard time remembering that the company still
mails out DVDs in small red packages, or the public relations disaster in 2011 that
had subscribers leaving in droves.
After sinking millions into original content and developing hit shows such as "House
of Cards," Netflix became the past year's best performer on the S&amp;P 500.
Netflix stock increased nearly 300 percent this past year, and now investors must
decide whether the media company's outperformance can carry into 2014.
Raymond James' Aaron Kessler told CNBC on Tuesday that the stock's current levels—
trading at around $365 on Tuesday—has priced in much of the company's upcoming good
news.
```

   As you can see, each sentence is delimited by return.

2. Word Tokenization & Stopwords Removal:

   Strategy adopted here is to tokenize words by space and set up a set of stopword with the help of nltk.corpus. After this step, all words are tokenized and stopwords are removed. You can review word_token.py for the source code. And output of this step is output_stopwords.txt.

```
REUTERS/China DailyHSBC China PMI fell 50.5 December, 50.8 previous month. A reading
50 indicates contraction. But right line expectations 50.5, flash PMI print. "The
moderation December's final HSBC China Manufacturing PMI mainly due slower output
growth," said Hongbin Qu, HSBC chief economist China, press release. "However, final
PMI sustained fifth above-50 reading row thanks steady increase new orders. The
recovering momentum since August 2013 continuing 2014, view. With inflation still
benign, expect current monetary fiscal policy remain place support growth." We got
```

All stop words have been removed and words are separated by space.

3. The next step is about stemming. The algorithm I used is porter in nltk python package. You can review the code in Stemming_.py. And result is output_of_stem.txt.

```
REUTERS/China DailyHSBC China PMI fell Decemb previou month read indic contract But
right line expect flash PMI print The moder Decemb final HSBC China Manufactur PMI
mainli due slower output growth said Hongbin HSBC chief economist China press releas
Howev final PMI sustain fifth above-50 read row thank steadi increas new order The
recov momentum sinc August 2013 continu 2014 view With inflat still benign expect
current monetari fiscal polici remain place support growth got China offici PMI data
Tuesday show manufactur fall Novemb The slowdown credit growth said weigh Chines
manufactur Here trajectori Chines manufactur HSBC/Markit Econom FREE AppDownload
Netflix user investor may hard time rememb compani still mail DVD small red packag
public relat disast 2011 subscrib leav drove After sink million origin content
develop hit show Hous Card Netflix becam past year best perform S&amp; 500 Netflix
stock increas nearli 300 percent past year investor must decid whether media compani
```

4. Extracting percentage:

   Since the percentages in the raw materials can be easily defined by regular expression. All details could be find in percentage.py. And the pattern about how to extract percentage could also be found there. Output of percentage could be find in output_of_percent.txt.

```
 50% 1.9% 0.9%  30% 15.3% 25.7% 19.3% 16.8%  10%  10% 3.5%  10%  19%  10%  128%  80%
0.9% 1.0% 1.0%  10% 4.8% 4.8% 2.3% 2.3% 2.6% 2.6%  2% 5.9% 1.8% 1.4% 0.3% 5.4% 1.7%
2.2% 1.8% 3.1% 11.9% 0.96% 0.85%  77%  100%  40% ~10%  40% -55%  50%  10% 0.6% 5.9%
0.9% 1.7%  16% 1.9% 16.6% 6.4% 13.4%  10% 3.6% 1.4% 7.8% 1.6% 1.5% 8.7% 5.6% 10.8%
15.7% 9.3%  10%  11%  53%  65% 7.1% 4.8% .75% 57.3% 56.5% 63.6% 1.0% 0.5% 0.4% 0.7%
15%  700%  10%  90%  10%  45%  400%  800% 6.3% 1.5%  80%  30% 15-20%  20% 1.5% 3.0%
16% 30-40%  30% 4.1% 7.7% 5.7% 8.4% 1.7% 4.3% 6.3% 1.5% 1.7% 3.1%  37%  25%  20%  20%
30% 4.1% 7.7% 7.9% 7.0% 21.5%  30% 2.6% 0.8% 0.9%  60%  30%  28%  55%  30% 1.2%  40%
```

5. POS tagging:

   In order to train the binary classifier to pick out all words which might be candidates of CEO names and companies' name. I have to finish the POS tagging first. Because all names are identified by tag = NNP. Nltk packages are used here to do the tagging job and we get a relatively nice result.

```
REUTERS/China__NNP DailyHSBC__NNP China__NNP PMI__NNP fell__VBD Decemb__NNP
previou__JJ month__NN read__VBD indic__JJ contract__NN But__CC right__JJ line__NN
expect__VBP flash__NN PMI__NNP print__NN The__DT moder__NN Decemb__NNP final__JJ
HSBC__NNP China__NNP Manufactur__NNP PMI__NNP mainli__NN due__JJ slower__JJR
output__NN growth__NN said__VBD Hongbin__NNP HSBC__NNP chief__JJ economist__NN
China__NNP press__NN releas__NN Howev__NNP final__JJ PMI__NNP sustain__NN fifth__JJ
above-50__JJ read__NN row__NN thank__VBD steadi__JJ increas__JJ new__JJ order__NN
The__DT recov__NN momentum__NN sinc__NN August__NNP 2013__CD continu__NN 2014__CD
view__NN With__IN inflat__NN still__RB benign__JJ expect__VBP current__JJ
monetari__JJ fiscal__JJ polici__NN remain__VBP place__JJ support__NN growth__NN
got__VBD China__NNP offici__MD PMI__NNP data__NNS Tuesday__NNP show__VBP
manufactur__JJ fall__NN Novemb__NNP The__DT slowdown__NN credit__NN growth__NN
```

6. Training SVM classifier:

   The reason why I choose SVM over Logistic regression is that I am not sure weather we can used a linear hyper plane to distinguish candidates from non-candidates. Considering the relatively slow convergence speed of Neural Network, SVM is the ideal supervised learning algorithm .

2. Feature selection of binary classifier.

a). I used 1/3 observations of raw data set as training set. It is not hard to find out patterns in ceo.csv and companies.csv. All CEO first names start with a uppercase letter and finished with a lowercase letter. The attributions of First Upper Letter and First Lower Letter are reasonable to build. Also, the family name of CEO also follow the pattern. And hence Second Upper Letter and Second

Lower Letter are also chosen as attributes. Also the length of name is also an important attribute. And POS attribute equals one if POS == NNP and 0 otherwise. I code every bigram up as a single observation, if the first word start with uppercase letter , column FirstUpperLetter is 1 , otherwise 0. Column SecondUpperLetter , FirstLowLetter,FirstUpperLetter follow same rules. Count the length of bigram as an another attribute. With respect to label of the training set. I set up a dictionary of CEO name according to ceo.csv. Observation which has same name in that dictionary will be labels as 1. Otherwise 0.

b).   Similar rules of Company name classifier are also adopted. But since there might be three of four words, I added ThirdUpper Letter and FourthUpperLetter. The accuracies of both model are high. Because of lot label_zero samples.

0.984180437954

	precision	recall	f1-score	support
0	0.99	0.99	0.99	894656
1	0.00	0.00	0.00	4800
avg / total	0.99	0.98	0.99	899456

0.97389199694

	precision	recall	f1-score	support
0	0.98	0.99	0.99	885402
1	0.00	0.00	0.00	14054
avg / total	0.97	0.97	0.97	899456

3. Dictionary trick.

Well, I review the results coming from the classifier. Unfortunately, my two classifiers failed to distinguish CEO name from Company name. Both classifiers just wrap them up. Therefore, I came up with an idea to solve this problem. I set up two dictionaries , one is about CEO name and another one is about company. And every time when I examine results from classifiers, I look up in those dictionaries. If it is the key of CEO dictionary, I assigned this observation to CEO names, otherwise companies' names. And the results are really good, you can check it at output_of_ceo.txt and output_of_company.txt.

```
Pete Wilson Russell Vic Barry Buffett Smith Sechin Woodman Virginia Jan Bernard Rosen
Simon Marsh Napier Internet Cooper White Karpeles Pershing McMahon Brown Twitch Jobs
Dunn Stein Brand Jordan Lozoya Katzenberg Zuckerberg Draghi Disney Weil Evan Fisher
Neither Day Joe Scharf Whitman Rogers Island Hancock Peter Michael Medvedev Safra
Networks Roberts Mulally John Nelson Jefferies Sony Art Ann Saunders Wolf Thomas
Parker Wynn Sam Schwartz Barra Sullivan Brent Cohen Ullman Rodriguez Spencer Sands
Paulson Allen Branson George Sandy Warren Einhorn Case Management From Strianese
Duncan Andrew Tesco Kravis Long Ellison Lee Horowitz Greenhaus Wendy Circle Jonathan
Jeffries Ross Musk LaSorda Mike Edelman Stade Tepper Rose Frank Holt Khodorkovsky
Fink Oberhelman William Dave Office Black Terry Cook Walsh Konheim Meyer Ketchum Jose
Benioff Brady Johnson Philippe Steinhafel Baker Bill Hanson David Bezos Yellen
Express Stanley Davis Fox Think Block First Reid Taylor Dean Clark Charney McDonald
Jeffery Kostin Layton Ballmer Chin Bain Ryan Graham Bass Moore Given Satya Group
Gross Robert Mara Nick Dalio Health Brooks Berman Shah Bernstein Horton Realty
Davidson Blankfein Hank Chanos Kerry Monday Adelson Byrne Mozilo Jamie Tsai Page
```

Figure 1 CEO name

```
Canada Autonomy Sun Residential Delphi Albuquerque Roche Leap Republic Asset
Citigroup Today Civeo Dodge Foundation Business Packing Standard Siemens Kellogg
Gazprom Western Zhejiang Odyssey Euro Grill Prudential Ventures Benzinga Alaska Yield
Yet Tech Lorillard Aviation Fixed Stryker Rand Emerging Nanex Huawei Global Tyco
DirecTV SumZero Nasdaq Defense Sands DoubleLine Sandy Chase Bausch Pharma Theranos
Andreessen Long State Interest Hillshire Jeffries Korea KPMG Paribas Ford Amazon
Hamilton Pepsi WhatsApp America Fort Family Motorola Mountains Freedom Budget Green
Cargill Ltd Habit Pimco PetSmart Warner CSX Merrill PepsiCo Chinese Energy Palantir
Resorts Associates Motor Home Inc Technology Audi Mountain Group Matrix City Silica
Three Nike Cantor Health Hill Perspectives AbbVie Australia Hub Hewlett-Packard
Cadillac Express Think Tiger Matters Visa Corp Peak Media Lockheed Land Boeing Intel
Oculus Americas Banking American Beverage Time Charter Financial Blizzard Citi Weibo
Actavis Point Pacific Grade Barclays Nikkei Nicola Nevada Panasonic Retirement IBM
Federal Corporation PayPal Reality Forbes Holding Digital Australian Sinopec AIG
Medtronic Agricole Eagle Lloyds Marketing Coke Dell Brewery Wayfair Chaori Dropbox
Swatch Athena Rental Industry Philips Television Hedgeye Heineken Nissan Bankers
Clorox Macquarie Viacom Mae Corporate Dow Raymond Uber Yukos Apple Valeant Community
One Celgene Cisco Bridgewater BHP Trust Priceline Journal Blue Barnes Blackstone
Smith Mutual Airbnb Covington Publicis Income White Edelman Smithfield Texas Shell
Scientific Verizon Exchange SAC Manitowoc Cup Xerox Brooks Lululemon Tesco World
```

Figure 2 Company name

4. Conclusion

Although two classifiers share relatively good accuracy. But they still failed to distinguish CEO names from Company names, which means I failed to find attributions which can tell me big difference between those two different kinds of names. However, I still get a fairly food result by looking up a dictionary. I think I should spend more time in finding good attribution. It is very important when it comes to data science.