Report

Jing Wan
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

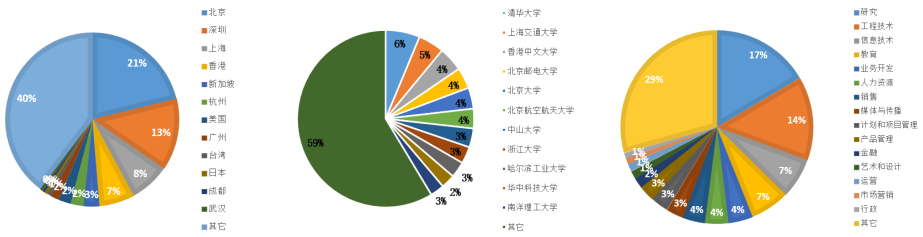
1 The first stage - basic information gathering

The main goal of the first phase is:

- Understand the basic information of the entire corporate executives, including academic background, academic achievements, and development process.
- The overall situation of obtaining the employees of Shangtang in LinkedIn China mainly includes the distribution of employees, the distribution of employees' colleges and universities, and the distribution of employees.

The main result of the first phase is:

1. The management team is generally from the top universities in the world, and has rich academic achievements and strong scientific research in the field of image recognition.
2. Shangtang Technology employees generally graduated from key universities around the world. The sample of employees from domestic 985 universities accounted for more than 41%, and the number of doctors exceeded 60. The personnel are mainly distributed in Beijing and Shenzhen. Researchers account for a large proportion, with researchers and engineers in the sample accounting for more than 31%.



(a) Employee area distribution. (b) Employee graduation college distribution. (c) Employee position distribution.

Figure 1: Basic information gathering.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 The second stage - crawler

Since the company information is constantly updated and there are many management personnel, it is obviously time-consuming and labor-intensive to collect information one by one at regular intervals. And the latest information update of the company is not timely, which may lead to wrong judgment of the company, so the goal of the second stage is: learning crawler technology. We attempt to capture relevant news of the company from the macro level, the industry level, the capital market, and the company level.

The implementation process is:

1. Determine the site segments to be crawled and classify them into four aspects: macro, industry, company, and capital market.The finalized websites include Securities Times, Hexun.com, Eastern Fortune.com, the financial sector, and Zhongcai.com, and finally decided to climb nearly 20 websites.
2. Based on BeautifulSoup framework, craw news from these static websites, Get the title, publish time.
3. Use multi-threading to allow crawlers of different web pages to execute concurrently to shorten crawling time.

```
# 证券时报网 -> 公司
def stcn_company(now, articles, titles):

    html = requests.get('http://company.stcn.com')
    soup = BeautifulSoup(html.content,'lxml')
    news_list = soup.find_all('ul', class_='news_list')[0].find_all('li')
    for news in news_list:
        title = news.p.a.contents[0].replace('\n','').replace('\r','').replace('\u2028','')
        title = re.sub('\s+', ' ', title, count=0, flags=0)
        if title in titles:
            continue
        time_list = news.p.next_sibling.next_sibling.next_sibling.next_sibling.contents
        time_str = time_list[0] + ' ' + time_list[1].contents[0]
        time_stamp = int(time.mktime(time.strptime(time_str, "%Y-%m-%d %H:%M")))
        if now - time_stamp > 3 * 24 * 60 * 60:
            break
        titles.append(title)
        article = Article(title, time_stamp, 'gongsi')
        articles.append(article)
```

Figure 2: Code (example).

3 The third stage - word segmentation and clustering

Due to the large number of titles taken out, the key information needs to be extracted and stored in an Excel table, updated once a day. The goal of this stage is to extract information from the crawled content, store the content in the excel table, update it daily, and form the word cloud of the recently updated day into a word cloud for visual display.

The implementation process is:

1. Identify the keywords you want to crawl.
2. Crawl once a day, word segmentation of the title taken every day
3. Using TF/IDF algorithm to count the five words with the highest frequency of words after adjustment, as key information. TF-IDF algorithm: The words that appear most frequently in the article are not necessarily keywords, such as the common stop words that do not make much sense to the article itself. IDF (Inverse Document Frequency) is inversely proportional to the commonality of a word. Multiply the word frequency (TF) and IDF to get the TF-IDF value of a word. The higher the importance of a word to an article, the greater its TF-IDF value, so the first several words are the keywords of the article.
4. Compare the key information extracted in the latest day with the key information of the previous day. If the extracted key information overlaps more than 60%, it means that there

is no information update on the current day. If it does not exceed 60%, the new information appears to be updated, and the extracted key information and time are updated in the excel form.

5. Visualize and display the word cloud for news of all keywords updated on the latest day.

```
keyword: 商汤
time: 2019/05/14 10:37
key infomation: 独角兽 旷视 影谱 15 AI
title: 商汤、旷视、影谱等15家AI企业入选一季度胡润独角兽指数
--------------------------------------
keyword: 人脸识别
time: 2019/05/06 21:16
key infomation: 骗保 人脸识别 欺诈 医保 北京市
title: 北京市医保局：通过人脸识别等持续打击欺诈骗保
```

Figure 3: Output in screen.

Figure 4: Wordcloud.

	A	B	C	D	E	F
1	商汤			人脸识别		
2	2019/5/14 10:37	独角兽 旷视 影谱 15 AI	商汤、旷视、影谱等15家AI企业入选一季度胡润独角兽指数	2019/5/6 21:16	骗保 人脸识别 欺诈 医保 北京市	北京市医保局：通过人脸识别等持续打击欺诈骗保

Figure 5: Output in excel.