



Clustering based on Vector Quantization and its applicable examples

20160742 강남웅

Contents



- I. Introduction
- II. Quantization
 - A. Scalar Quantization
 - B. Vector Quantization
- III. LBG Algorithm
- IV. K-means Clustering
- V. Application
 - A. Image Compression
 - B. Document Clustering
- VI. Reference

Introduction



Data compression has two main type.

1. Lossless Compression - mainly used for text data compression
 - a. Hoffman Coding
 - b. Lempel-Ziv Algorithms
2. Lossy Compression - mainly used for image / voice data compression
 - a. Quantization
 - b. Transform Coding
 - c. Wavelet Compression

“ This presentation mainly focuses on VQ (vector Quantization) “

Quantization



Definition : Process of mapping input values from a large set to output values in a countable smaller set
(Reference - Wikipedia)

“ Two stage of Quantizer : Encoder mapping / Decoder mapping ”

Two types of Quantization :

- Scalar Quantization
- Vector Quantization

Scalar Quantization

Definition : Set S comes from a total order and total order is broken up into regions that map onto the elements of smaller set S'

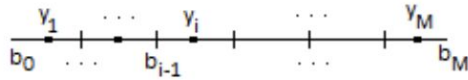
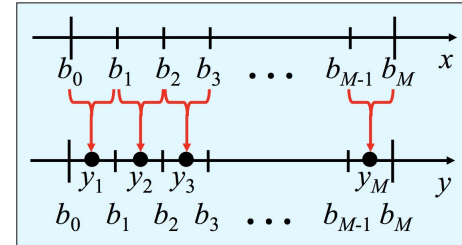


Figure 1: Uniform quantization



- Lloyd-Max Algorithm

Vector Quantization



Definition : Mapping multi-dimensional space S into a smaller message space of S'

selecting a set of representatives from input space



map to closest representative

Main objective : how to select representatives (Center of cluster / Centroid)

=> Therefore, we can view VQ as clustering

LBG Algorithm



Pseudocode

1. Input training vectors $S = \{\mathbf{x}_i \in R^d \mid i = 1, 2, \dots, n\}$.
2. Initiate a codebook $C = \{\mathbf{c}_j \in R^d \mid j = 1, 2, \dots, K\}$.
3. Set $D_0 = 0$ and let $k = 0$.
4. Classify the n training vectors into K clusters according to $\mathbf{x}_i \in S_q$ if $\|\mathbf{x}_i - \mathbf{c}_q\|_p \leq \|\mathbf{x}_i - \mathbf{c}_j\|_p$ for $j \neq q$.
5. Update cluster centers \mathbf{c}_j , $j = 1, 2, \dots, K$ by $\mathbf{c}_j = \frac{1}{|S_j|} \sum_{\mathbf{x}_i \in S_j} \mathbf{x}_i$.
6. Set $k \leftarrow k + 1$ and compute the distortion $D_k = \sum_{j=1}^K \sum_{\mathbf{x}_i \in S_j} \|\mathbf{x}_i - \mathbf{c}_j\|_p$.
7. If $(D_{k-1} - D_k)/D_k > \epsilon$ (*a small number*), repeat steps 4 ~ 6.

LBG Algorithm

Implementation

```
def lbg(data, K):  
    num_of_data = len(data)  
    dim = len(data[0])  
  
    centroids = []  
    while len(centroids) != K:  
        for i in range(K):  
            centroids.append( [ random.randint(0,255) for j in range(dim) ] )  
  
    error = []  
    distortion = 9999  
    k = 0  
    while True:  
        codeword, S, err = get_cluster(K, centroids, data)  
        error.append(err)  
  
        if k == 0:  
            distortion = error[-1]  
        else:  
            distortion = (error[-2] - err) / err
```


LBG Algorithm

Implementation

```
if distortion < 0.1:
    break
else:
    new_centroids = []
    for _cds in S:
        _tmp = [0] * dim
        _len = len(_cds)
        if _len != 0:
            for _c in _cds:
                for i in range(dim):
                    _tmp[i] += _c[i]
            for i in range(dim):
                _tmp[i] = _tmp[i] / _len

            new_centroids.append( _tmp )
        else:
            new_centroids.append( [ random.randint(0,255) for j in range(dim) ] )

    centroids = new_centroids
    k += 1
return centroids, S, codeword
```

LBG Algorithm



Implementation

```
cent, S, codeword = lbg(data, 10)
```

```
=====
error : 9276468.399890926
=====
=====
error : 0.8945696015482465
=====
=====
error : 0.6126551460839871
=====
=====
error : 0.09930588679143644
=====
```

K means Clustering



Definition : A method of vector quantization that aims to partition n observation into k clusters.
(Reference : Wikipedia)

- Based on EM Algorithm
 - E step :
Split input set by distance between input data and centroid of a cluster
 - M Step:
Update centroid to center of each cluster
- One of the most famous clustering algorithm
- Lightweight Algorithm

Application - Data Compression (LBG)

Example) cat.jpg



Original Image : 26,334byte



Application - Data Compression (LBG)

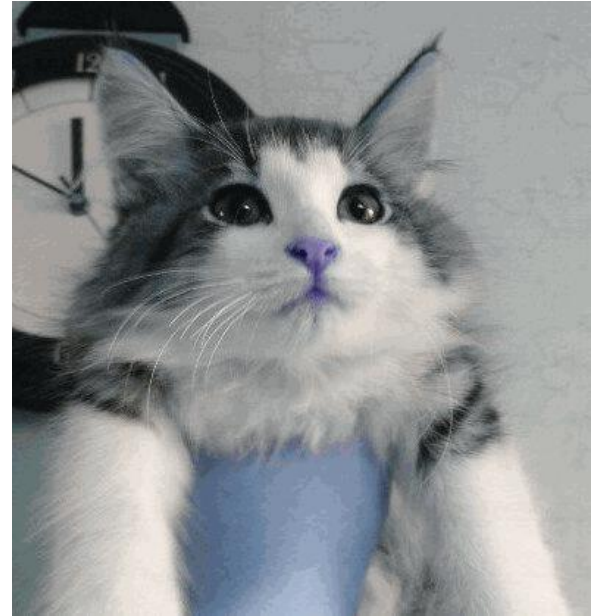
Original -> 200 clusters

Read Image & run LBG Algorithm

```
import cv2
img_1 = cv2.imread('cat.jpg')

data = []
for i in range(img_1.shape[0]):
    for j in range(img_1.shape[1]):
        data.append( img_1[i][j].tolist() )
```

```
cent, S, codeword = lbg(data, 200)
```



Result Image

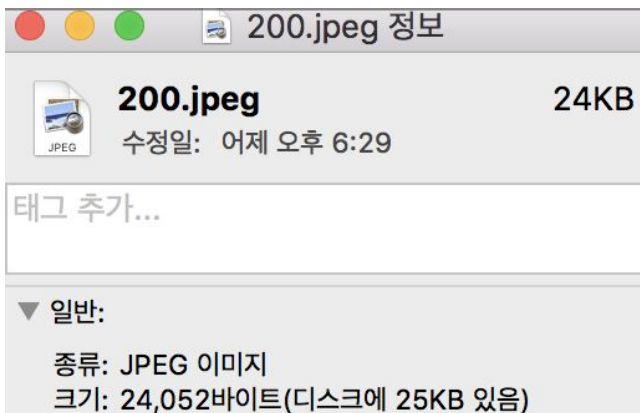
Application - Data Compression (LBG)

200 cluster result



Data size : 26,334byte ->24,052byte

Compressed rate : 91.3%



Application - Data Compression (LBG)

Original -> 100 clusters

Read Image & run LBG Algorithm

```
import cv2
img_1 = cv2.imread('cat.jpg')

data = []
for i in range(img_1.shape[0]):
    for j in range(img_1.shape[1]):
        data.append( img_1[i][j].tolist() )
```

```
cent, S, codeword = lbg(data, 100)
```



Result Image

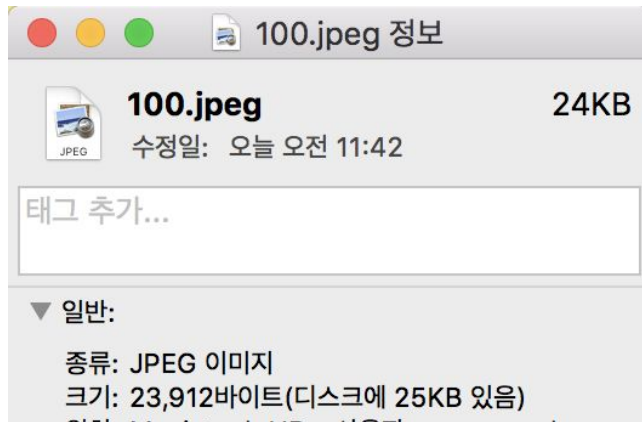
Application - Data Compression (LBG)

100 cluster result



Data size : 26,334byte ->23,912 byte

Compressed rate : 90.8%



Application - Data Compression (LBG)

Other Results



20 cluster



10 clusters



2 clusters

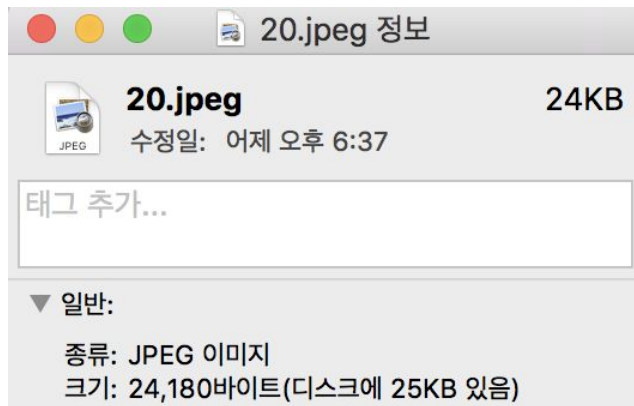
Application - Data Compression (LBG)

20 cluster result



Data size : 26,334byte ->24,180 byte

Compressed rate : 91.8%



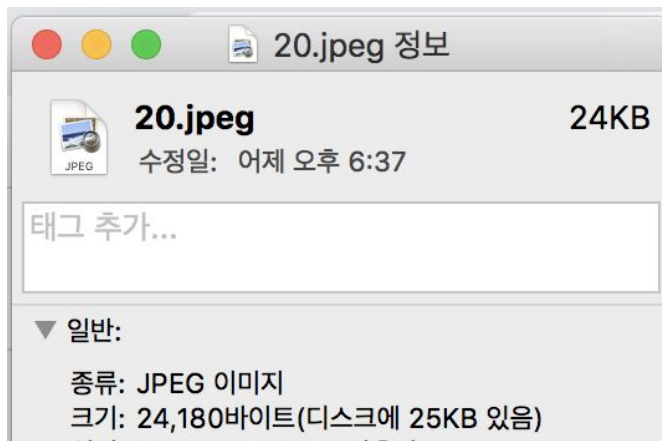
Application - Data Compression (LBG)

10 cluster result



Data size : 26,334byte ->24,180 byte

Compressed rate : 91.8%



Application - Data Compression (LBG)

2 cluster result



Data size : 26,334byte -> 16,076 byte

Compressed rate : 61.0%



Discussion



Possible Optimization

- Non - randomized Initialization for Centroids
 - An Efficient Codebook Initialization approach for LBG Algorithm
 - <https://arxiv.org/pdf/1109.0090.pdf>

Application - Document Clustering

Main goal : Given 1K of news articles, make clusters of related news articles

Input Data : Naver news articles

(Main Category - Economy / Sub Category - Finance)

4 Main Stages :

Data Crawling -> Data Preprocessing -> Vectorize -> Clustering



The screenshot shows the Naver News Finance section. The top navigation bar includes '뉴스' (News), 'TV연예' (TV Entertainment), '스포츠' (Sports), '뉴스스탠드' (News Stand), and '날씨' (Weather). Below this, a sub-navigation bar lists various categories: '뉴스' (News), '속보' (Breaking News), '정치' (Politics), '경제' (Economy), '사회' (Society), '생활/문화' (Life/Culture), '세계' (World), 'IT/과학' (IT/Science), '오피니언' (Opinion), '포토' (Photo), 'TV' (TV), and '랭킹뉴스' (Ranking News). The main content area is titled '11.30 (월) 헤드라인 뉴스' (11.30 (Mon) Headline News) and features a list of news articles. The left sidebar contains a '증권' (Certificate) section with links to '금융' (Finance), '증권' (Securities), '산업/재계' (Industry/Business), '중기/벤처' (Mid-size/Venture), '부동산' (Real Estate), '글로벌 경제' (Global Economy), '생활경제' (Life Economy), '경제 일반' (General Economy), and '속보' (Breaking News). The main content area displays three news articles: 1. '누구 위한 IPO제도 개편?'...개미기관 모두 불안 (Who is the IPO system reform for?...All the little institutions are anxious), 2. '사학연금, 조선대학교와 산학협력 증진 위한 업무협약 체결' (Private school pension, Chosun University and industry-academia cooperation promotion business cooperation agreement signed), and 3. '지존심 구진 교촌치킨, 나홀로 근무박질...반동할까' (Jeeon Sim Gu-jin, Gochon Chicken, alone working hard...Will it be a counter-revolution?).

Application - Document Clustering

1. Data Crawling - Target Example / Code

매일경제

토스증권 증권업 본인가 획득

기사입력 2020.11.18. 오후 4:17 | 기사원문 | 스크랩 | 본문듣기 · 설정

공감 댓글

요약본가

금융위 정례회의 통과

토스증권이 국내 첫 모바일 전문 증권사로 공식 출범한다.

금융위원회는 18일 정례회의에서 토스증권 준비법인에 대해 증권업 진출을 위한 투자중개업 본인가 안을 통과시켰다. 토스증권은 이번 인가를 통해 내년초 영업을 개시할 계획이다.

토스증권은 핀테크업체 비바리퍼블리카의 계열사로, 자본금 340억원을 마련하고 올해 직원 80명을 고용하는 등 증권업 진출을 준비하고 있다.

토스증권은 국내 주식 중개를 시작으로, 향후 해외주식 중개, 집합투자증권(펀드) 판매로 서비스를 확장할 계획이다. 토스증권은 2030밀레니얼 세대를 타겟으로 MTS(모바일 트레이딩 시스템)를 중심으로, 차별화된 투자정보를 제공한다는 방침이다.

박재민 토스증권 대표는 "투자 입문자의 시각에서 MTS의 모든 기능을 설계하고, 메뉴의 구성이나 명칭, 투자 정보의 탐색 등 주요 서비스를 완전히 새롭게 구성했기 때문에, 기존 증권사의 MTS가 복잡

```
def get_article(article):
    res_article = requests.get(article.select('a')[0]['href'], headers=headers)
    r_article = BeautifulSoup(res_article.text, 'lxml')

    title = r_article.select('h3#articleTitle')[0].get_text().strip()

    for elem in r_article.select('div#articleBodyContents')[0].select('span.end_photo_org'):
        elem.decompose()
    for elem in r_article.select('div#articleBodyContents')[0].select('script'):
        elem.decompose()
    for elem in r_article.select('div#articleBodyContents')[0].select('a'):
        elem.decompose()

    body = r_article.select('div#articleBodyContents')[0].get_text().strip()

    return [title, body]
```

Application - Document Clustering

1. Data Crawling - Result

A	
title	
금융당국 인가 없이 비상장주식 투자 불법자문 26명, 검찰 송치	(서울=뉴스1) 박종홍 기자 = 금융당국의 인가를 받지
코픽스 하락에도...주담대 금리 높아져	국민·농협은행 등 '우대' 없애주담대 금리 되레 소폭
산은, 한진칼 사외이사 3인 지명 등 경영권 적극 개입	주요 경영사안 사전 협의윤리경영특·경영평가특 설치
'옵티머스 로비 의혹' 신 회장 구속..."주요 혐의 소명됐다"	'핵심 로비스트' 중 한 명으로 지목법원 "도망·증거인멸
옵티머스 핵심 로비스트 前연예기획사 대표 구속영장 발부	법원 "혐의 사실 소명...도망·증거 인멸 염려 있어"(서울=
국민연금, KB금융 우리사주조합 추천 사외이사 '반대'	"장기적 주주까지 중대 불확실" 윤종규·하인 사내이
7대의무 부여 받은 한진칼...위반시 5000억 위약금(상보)	[머니투데이 박광범 기자] KDB산업은행(이하 산은)
카카오뱅크, 홍콩 PEF서 2500억 투자 유치	[한국경제TV 김보미 기자]카카오뱅크가 17일 오후 (
산업은행, 화재피해 장애인시설에 2000만원 후원	산업은행은 지난 16일 'KDB 따뜻한 동행' 40호 후원
KB국민은행, 독립유공자 후손 지원금 4억 전달	KB국민은행은 대한적십자사에 31억원 101주년 기
NH투자증권, 11월 '100세시대 아카데미' 유튜브 세미나 개최	25일 오후 3시30분부터 90분간 유튜브로 진행내년
우리은행, 'WON금융인중서' 출시	[파이낸셜뉴스] 우리은행은 금융권 최초로 클라우
카카오, 영커에쿼티파트너스에서 2500억원 추가투자유치	[파이낸셜뉴스] 카카오뱅크는 사모투자펀드인 영
산은, 한진칼에 7대의무 부과...위반시 5000억 위약금	[머니투데이 박광범 기자] KDB산업은행(이하 산은)
"대출 막히기 전에" 주담대사외이 신용대출 2배 폭증	13일 규제 발표 후 비대면 금융30일 이전 '막자' 고
오른병집 참여 조건에 '정보 제공 기관' 추가 가락	카드사 공식적인 참여 방법 생김핀테크, 선불충전금
은행연합회장 후보군 확정...민병두·김광수·신상훈 '3파전'	회주영, 정·관·민 출신 7명 발표민병두, '3선 국회의
신한금융, 2030세대 자산감 키우기 캠페인 '선로	신한금융그룹은 지난 3개월간 밀레니얼 세대들을
카카오페이 '내 대출 한도', 제휴 금융사 30곳 돌파	이제 카카오페이 안에서 30개 금융사의 신용 대출
삼상생명도 진출...드라워진 알리보험 시장	오늘부터 '알리중신보험' 선보여국내보험사 잇단 진
롯데손보, 업계 첫 소방관 대상 전용보험 상품 출시	롯데손해보험(대표 최원진)은 업계 최초로 소방관들
産權, 한진칼 이사 선임권...産權 "특에 아니다" 긴급진화	주요 경영사항 등의 의무화산업은행이 한진칼에 야
"산은 동의 없이 경영 못한다"...한진 폭출 침 체권단	산은·한진칼과 투자합의서 체결,7대 의무사항 담
경영평가 부진 땀 해임 '조강수'...산은에 담보 잡힌 조원태	(산은, 한진칼에 7대 의무 부과)윤리경영·평가위 설치
조원태 백기사 시비 불복...産權 "특에 아니다" 긴급진화	産權 7대 의무사항 공개경영권 견제조항 내걸었지만
산은·한진칼, 투자합의서 체결,위반시 위약금 5000억	[파이낸셜뉴스] 산업은행은 17일 한진칼과 8000억
미래에셋생명, '비밀증치해보험' 확인 특약 출시	미래에셋생명은 업계 최초 비보험자에게 치해보
신용대출 속도 증단...염골막자' 띄워보	11·13 대출 규제 이후신용대출 신규 신청 폭증하자
주담대기준 코픽스 떨어졌대 국민·농협銀, 금리 되레 올려	"우대금리 일괄 조정 때문"변동형 주택담보대출 금

```
import pandas as pd
```

```
df = pd.read_excel('information_theory_news_data.xlsx')
```

```
df.to_excel('tmp1.xlsx', index=None)
```

```
df.head()
```

	title	body	preprocessed
0	금융당국 인가 없이 비상장주식 투자 불법자문 26명, 검찰 송치	(서울=뉴스1) 박종홍 기자 = 금융당국의 인가를 받지 않고 수천억원 대 비상장 주...	금융당국의 인가를 받지 않고 수천억원 대 비상장 주식 투자를 자문해온 업체 대표 등...
1	코픽스 하락에도...주담대 금리 높아져	국민·농협은행 등 '우대' 없애주담대 금리 되레 소폭 올려지난달 10개월 만에 반등...	국민 농협은행 등 우대 없애주담대 금리 되레 소폭 올려지난달 10개월 만에 반등...
2	산은, 한진칼 사외이사 3인 지명 등 경영권 적극 개입	주요 경영사안 사전 협의윤리경영특·경영평가특 설치 등7개 의무 위반엔 5000억 위...	주요 경영사안 사전 협의윤리경영특·경영평가특 설치 등7개 의무 위반엔 5000억 위...
3	'옵티머스 로비 의혹' 신 회장 구속..."주요 혐의 소명됐다"	'핵심 로비스트' 중 한 명으로 지목법원 "도망·증거인멸 염려 인정돼"검찰, 정·관...	핵심 로비스트 중 한 명으로 지목법원 도망 증거인멸 염려 인정돼 검찰 정관계...
4	옵티머스 핵심 로비스트 前연예기획사 대표 구속영장 발부	법원 "혐의 사실 소명...도망·증거 인멸 염려 있어"(서울=뉴스1) 윤수희 기자 = ...	법원 혐의 사실 소명 도망 증거 인멸 염려 있어 옵티머스 자산운용 펀드 사기...

Application - Document Clustering

2. Data Preprocessing - Remove unrelated text (Email address)

이러스가 페인트에 붙으면 30분 경과 후부터 바이러스 감소효과를 보이기 시작해 24시간 내 99.9%가 사멸되는 것으로 나타났다.

삼화페인트 항바이러스 페인트는 12월 중 출시되며 해썹인증 제조업체인 식자재 전문 브랜드 화미에 적용 예정이다.

khj91@fnnews.com 김현정 기자

```
df.iloc[0][1]
```

'문재인 대통령이 미국 연방 의회에서 역대 최대인 4명의 한국계 의원이 당선된 데 대해 17일 "무척 고무적이다. 앞으로 한미관계 발전을 위해서도 함께 협력해 나가길 기대한다"고 밝혔다. 문 대통령은 이날 소셜네트워크서비스(SNS)에 올린 글에서 "미연방 의회 한국계 의원들의 당선을 축하한다"며 이같이 전했다. 문 대통령은 이들의 당선 소식에 대해 "기쁘고 유쾌하다"며 "이 분들은 '영옥', '은주', '순자' 같은 정겨운 이름을 갖고 있다"고 했다. 그러면서 영 김(한국명 김영옥·공화당), 미셸 박 스틸(한국명 박은주·공화당), 메릴린 스트릭랜드(한국명 순자·민주당), 앤디 김(민주당) 당선인을 일일이 열거했다. 문 대통령은 영 김 당선인에 대해 "한인 방송 진행자로 활약하며 한인사회와 미 주류사회의 가교역할을 해왔다"고 했다. 미셸 박 스틸 당선인을 향해서는 "청소년 보호에 각별한 애정을 가지고 지역 커뮤니티 현안에 높은 관심을 보여왔다"고 했고, 메릴린 스트릭랜드 당선인에 대해서는 "시애틀 상공회의소 회장을 역임한 경제전문가"라고 했다. 재선에 성공한 앤디 김 의원에게는 "한국전 종전선언 촉구 결의안 발의 등 그동안 한반도 평화를 위해 누구보다 활발한 의정활동을 펼쳐 왔다"고 평가했다. 문 대통령은 "정겨운 우리 이름들이 더욱 근사하게 느껴진다"며 "무엇보다 이 분들이 계셔서 미국의 우리 한인들이 든든할 것"이라고 했다. 박효목기자 tree624@donga.com© 동아일보 & donga.com, 무단 전재 및 재배포 금지'

Application - Document Clustering

2. Data Preprocessing - Remove unrelated text (Pattern I)

“ xxx 기자 = ”

	title	body
0	금융당국 인가 없이 비상장주식 투자 불법자문 26명, 검찰 송치	(서울=뉴스1) 박종홍 기자 = 금융당국의 인가를 받지 않고 수천억원 대 비상장 주...
987	[속보] 코스피 2년6개월만에 2500 돌파	(서울=뉴스1) 전민 기자 = 장중 2018년5월3일 이후 처음min785@news...
4	옵티머스 핵심 로비스트 前연예기획사 대표 구속영장 발부	법원 "혐의 사실 소명...도망·증거 인멸 염려 있어"(서울=뉴스1) 윤수희 기자 = ...

Application - Document Clustering

2. Data Preprocessing - Remove unrelated text (Pattern II)

“ [~~] Related sentences.... ” or “ ~~ SNS(Social Network Services) ~~ ”

금융당국 인가 없이 비상장주식 투자 불법자문 26명, 검찰 송치

(서울=뉴스1) 박종홍 기자 = 금융당국의 인가를 받지 않고 수천억원 대 비상장 주...

[빅데이터MSI]시장심리 톱5, 삼성전자·대한항공·현대차·CJ·LG유플러스

위스트5, 현대제철·삼성물산·대우건설·쌍용차·LS [서울=뉴시스] 이승주 기자 = ...

상호협력을 통한 동반 성장을 위한 업무협약(MOU)을 체결했다고 17일 밝혔다.
'고기능 다통화 외화 자동입출금기(ATM)'를

Application - Document Clustering



2. Data Preprocessing - Remove Special Characters / Copyrights

"...모토는 같다"며 "공동의 목표를 바탕으로 마케팅과
r저작권자 © 서울경제, 무단 전재 및 재배포 금지'

△사망, 후유장애 △중증화상·부식진단

14~15일에는 온라인 비대면

Application - Document Clustering

2. Data Preprocessing - Code

- Email
- Remove unrelated text
- Remove Special Characters / Copyrights
- Text inside Parenthesis

```
def preprocess(a):
    target = a['body']

    # remove email address
    _email_list = re.findall("([a-zA-Z0-9_+~]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-\.]+)", target)
    if _email_list:
        target = target.split(_email_list[0])[0]

    # diversify 0.2 vs ~(sent). (sent)~ by replacing '.' to '. '
    result = ''
    for idx in range(len(target)):
        if target[idx] == '.':
            if not target[idx-1].isdigit() or not target[idx+1].isdigit():
                result += '. '
            else:
                result += '.'
        else:
            result += target[idx]

    target = result

    # remove pattern I - " XXX 기자 = 금융당국.... "
    target = re.sub(r'[가-힣]+ 기자\ =', '', target)
    # remove stuffs in parenthesis
    target = re.sub(r'\([([^\)]+)\)', '', target) # []
    target = re.sub(r'\([([^\)]+)\)', '', target) # ()
    # remove all special characters
    target = re.sub(r'[^a-zA-Z0-9가-힣-~\.\s]', '', target).strip()

    return target
```


Application - Document Clustering



3. TF-IDF : Frequently used statistical methodology for text analysis

$$\text{tf}(t, d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

Application - Document Clustering

3. TF-IDF : Frequently used statistical methodology for text analysis

1. I love dogs.

2. I hate dogs and knitting.

3. Knitting is my hobby and my passion.



	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Application - Document Clustering

3. TF-IDF - Code

Apply TF-IDF to “Title + Key sentence”

	title	key_sentence
0	금융당국 인가 없이 비상장주식 투자 불법자문 26명, 검찰 송치	금융당국의 인가를 받지 않고 수천억원 대 비상장 주식 투자를 자문해온 업체 대표 등...
1	코픽스 하락에도...주담대 금리 높아져	그런데 국민 농협 등 일부 은행은 코픽스 하락에도 변동형 주담대 금리를 올렸다
2	산은, 한진칼 사외이사 3인 지명 등 경영권 적극 개입	한진칼은 산은이 지명하는 사외이사 3인 및 감사위원회 위원을 선임해야 한다
3	'옵티머스 로비 의혹' 신 회장 구속..."주요 혐의 소명됐다"	구속 심사에 앞서 법원에 출석한 신씨는 로비 혐의는 부인하는 입장인지 옵티머...
4	옵티머스 핵심 로비스트 前연예기획사 대표 구속영장 발부	법원 혐의 사실 소명 도망 증거 인멸 염려 있어 옵티머스 자산운용 펀드 사기와...

```
corpus = []
for doc in df_1.itertuples():
    title = doc[-2]
    key = doc[-1]

    pre_title = re.sub(r'^a-zA-Z0-9가-힣-籲\ ]', ' ', title).strip()
    pre_key = re.sub(r'^a-zA-Z0-9가-힣-籲\ ]', ' ', key).strip()
    print( pre_title + ' ' + pre_key)
    print( '*' * 20 )
    corpus.append( pre_title + ' ' + pre_key)
```



```
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)
```

Application - Document Clustering

4. Clustering - K- means Clustering

```
from sklearn.cluster import KMeans
import numpy as np

num_of_clusters = 50

#Implement k-means clustering to form k clusters
kmeans = KMeans(n_clusters=num_of_clusters)
kmeans.fit(data)
```

```
cat_1 = kmeans.cluster_centers_[c_idx]
```

```
vectorizer.inverse_transform( cat_1 )
```



Key 20 words of cluster 5 :

['우리은행',
'WON금융인증서',
'출시',
'우리은행은',
'금융권',
'최초로',
'클라우드',
'기반의',
'다양한',
'기관에서',
'사용',
'가능한',
'를',
'출시했다고',
'17일',
'밝혔다',
'우리銀',
'원금융인증서',
'공공기관',
'어디서든']

Application - Document Clustering

Result - 50 cluster

Top 20 words of cluster 0 :

```
[array(['kb금융', 'kb금융지주', '경영에', '계열사', '고문', '사내이사', '산은', '선임', '않기로',  
       '우리사주', '윤종규', '이명희', '정석기업', '조현민', '찬성', '참여하지', '한진그룹', '항공',  
       '했다', '허인'],  
      dtype='<U16')]
```

Top 20 frequent word in cluster 0 data

```
['산은', '동의', '없인', '경영', '못한다', '한진', '목줄', '권', '채권단', '이들은', '항공', '관련', '계열사', '경영에', '참여하지', '않는  
다', '국민연금', 'KB금융', '우리사주', '이사제안']
```

10 article of given cluster :

"산은 동의 없인 경영 못한다"..한진 목줄 권 채권단
국민연금, KB금융 우리사주 이사제안 '반대'...윤종규·허인 찬성
국민연금, KB금융 우리사주 제안 사외이사 선임 반대..."윤종규·허인 사내이사 찬성"
국민연금, KB금융 윤종규·허인 찬성...우리사주 선임안 '반대표'
국민연금, '우리사주조합 추천' KB금융지주 사외이사 선임 반대
[일문일답] 산은 "조현민·이명희, 항공 계열사 경영참여 안 해"
산은 "조현민·이명희 등 한진그룹 일가, 항공계열사 경영참여 않기로"
산은 "한진 오너 일가, 항공 계열사 경영 참여 않기로"
산은 "조현민·이명희, 항공 계열사 경영 참여 않는다"
[속보] 산은 "조현민, 이명희 등 계열사 일가도 윤리감독에 참여"

Application - Document Clustering

Result - 50 cluster

Top 20 words of cluster 3 :

```
[array(['60만명이', '것이라고', '경우', '낮아질', '내렸지만', '밀려날', '밖으로', '법정', '어려워져',  
       '우려', '인하로', '전망했다', '제도권', '주택시장의', '최근', '한계차주', '한국대부금융협회는',  
       '혼합형'],  
      dtype='<U16')]
```

Top 20 frequent word in cluster 3 data

```
['머니', '컨설팅', '혼합형', '펀드로', '위험', '낮추고', '수익', '높이기', '펀드의', '경우', '펀드', '자체적으로', '자산', '배분', '절차  
를', '밟게', '돼', '변동성을', '완화하고', '안정적인']
```

10 article of given cluster :

[머니 컨설팅]혼합형 펀드로 위험 낮추고 수익 높이기

매출 타격 `등록대부업` 붕괴 가속화 우려

[툏아보기]주택시장의 KF94, 분할상환 전세대출을 아시나요

법정최고금리 인하로 등록대부업 붕괴 가속화 우려

법정 최고금리는 내렸지만...한계차주 대출 더 어려워져

법정 최고금리 내렸지만...한계차주 대출 더 어려워져

Application - Document Clustering

Result - 50 cluster

Top 20 words of cluster 6 :

```
[array(['10월', '87', '금리', '기준', '기준이', '다시', '되는', '만에', '소폭', '신규', '은행',  
       '일부', '주담대', '주택담보대출', '취급액', '코픽스', '코픽스가', '하락', '하락세', '한달만에'],  
      dtype='<U16')]
```

Top 20 frequent word in cluster 6 data

```
['코픽스', '하락에도', '주담대', '금리', '높아져', '그런데', '국민', '농협', '등', '일부', '은행은', '변동형', '금리를', '올렸다', '주담대기  
준', '떨어졌는데', '농협銀', '되레', '올려', '우대금리']
```

10 article of given cluster :

코픽스 하락에도...주담대 금리 높아져

주담대기준 코픽스 떨어졌는데 국민·농협銀, 금리 되레 올려

코픽스 하락에도...일부 은행 '주담대' 금리 올라

코픽스 하락했는데...일부 은행 주택담보대출 금리 오히려 올라

코픽스 한달만에 하락... 은행 주담대 금리는 '역주행'

코픽스 하락 전환에도...일부 은행 주담대 금리 오른다(종합)

은행 코픽스 다시 하락했지만...주담대 금리 오르는 '역주행'(종합)

신규 취급 코픽스 한달만에 하락...주담대 금리 내린다(종합)

주담대 금리 내려간다... 신규 취급액 코픽스 한달만에 하락

코픽스, 한 달 만에 다시 하락세...10월 신규 취급액 0.87%

.....

Discussion



Possible Optimization / Improvement

- More precise preprocessing methods
 - Stop words (그런, 이를, ..)
 - Auxiliary words (-를, -을)
 - Abbreviations (SNS, ATM, ...)
- New standard for Key sentence extraction
 - BLEU Score
- Effective Clustering Algorithms
 - kNN Algorithm
 - Neural Network

Reference



Wikipedia

[https://en.wikipedia.org/wiki/Quantization \(signal processing\)](https://en.wikipedia.org/wiki/Quantization_(signal_processing))

<https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

Vector Quantization

<http://people.ece.umn.edu/~arya/EE5585/lecture13.pdf>

[http://www.ws.binghamton.edu/fowler/fowler%20personal%20page/EE523_files/Ch_09_1%20SQ%20Overview%20\(PPT\).pdf](http://www.ws.binghamton.edu/fowler/fowler%20personal%20page/EE523_files/Ch_09_1%20SQ%20Overview%20(PPT).pdf)

LBG Algorithm

<http://www.cs.nthu.edu.tw/~cchen/CS4520/Notes/LBG.pdf>

Reference



Image Compression

<https://towardsdatascience.com/cluster-based-image-segmentation-python-80a295f4f3a2>

<https://towardsdatascience.com/image-compression-using-k-means-clustering-aa0c91bb0eeb>

<https://www.naun.org/main/NAUN/mcs/2002-109.pdf>

<https://www.kdnuggets.com/2019/08/introduction-image-segmentation-k-means-clustering.html>

<https://twlab.tistory.com/19>

K-means Clustering

<https://ratsgo.github.io/machine%20learning/2017/04/19/KC/>

Reference



Document Clustering

<https://konlpy-ko.readthedocs.io/ko/v0.4.3/>

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

<https://www.tutorialspoint.com/Extracting-email-addresses-using-regular-expressions-in-Python>



Clustering based on Vector Quantization and its applicable examples

Codes Available at

https://github.com/treeK-u/MAS370_Information_Theory