Automated Resume Screening

COMP 4750: Natural Language Processing

Shawon Ibn Kamal, 201761376

In [1]:
```
from os import path
from glob import glob


from pdfminer.high_level import extract_text

import nltk
from nltk.corpus import stopwords
import re
import subprocess

import pandas as pd
import numpy as np

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics.pairwise import euclidean_distances
```

Part 1: Parsing

Read resume pdf

In [2]:
```
mypath = "resumes-list"

def find_ext(dr, ext):
    return glob(path.join(dr,"*.{}".format(ext)))

resumepaths = find_ext(mypath, "pdf")

df = pd.DataFrame (resumepaths, columns = ['path'])


df['text'] = df['path'].apply(lambda x: extract_text(x))

df.head()
```

Out[2]:

	path	text
0	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...
1	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...
2	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...
3	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...
4	resumes-list/data-scientist-resume-example.pdf	KANDICE LOUDOR\n\nDATA SCIENTIST\n\nCONTACT\n\...

Retrieve candidate name

In [3]:
```
def extract_names(txt):
    person_names = []

    for sent in nltk.sent_tokenize(txt):
        for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(ser
            if hasattr(chunk, 'label') and chunk.label() == 'PERSON':
                person_names.append(
                    ' '.join(chunk_leave[0] for chunk_leave in chunk.l
                )

    return person_names

df['name'] = df.text.apply(lambda x: extract_names(x)[0])

df.head()
```

Out[3]:

	path	text	name
0	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS
1	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...	Stephen
2	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...	Ashley
3	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...	Stephen
4	resumes-list/data-scientist-resume-example.pdf	KANDICE LOUDOR\n\nDATA SCIENTIST\n\nCONTACT\n\...	Github

Extract phone-number

In [4]:
```
phone_regex = re.compile(r'[\+\(]?[1-9][0-9 .\-\(\)]{8,}[0-9]')

def extract_phone_number(resume_text):
    phone = re.findall(phone_regex, resume_text)

    if phone:
        number = ''.join(phone[0])

        if resume_text.find(number) >= 0 and len(number) < 16:
            return number
    return None

df['phone'] = df.text.apply(lambda x: extract_phone_number(x))

df.head()
```

Out[4]:

	path	text	name	phone
0	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS	(123) 456-7890
1	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...	Stephen	(123) 456-7890
2	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...	Ashley	(123) 456-7890
3	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...	Stephen	(123) 456-7890
4	resumes-list/data-scientist-resume-example.pdf	KANDICE LOUDOR\n\nDATA SCIENTIST\n\nCONTACT\n\...	Github	(123) 456-7890

Extract email

In [5]:
```
email_regex = re.compile(r'[a-z0-9\.\-+_]+@[a-z0-9\.\-+_]+\.[a-z]+')
def extract_emails(resume_text):
    return re.findall(email_regex, resume_text)

df['email'] = df.text.apply(lambda x: extract_emails(x))

df.head()
```

Out[5]:

	path	text	name	phone	email
0	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS	(123) 456-7890	[justin.green11@gmail.com]
1	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...	Stephen	(123) 456-7890	[stephen@beamjobs.com]
2	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...	Ashley	(123) 456-7890	[ashley.do@gmail.com]
3	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...	Stephen	(123) 456-7890	[stephen@beamjobs.com]
4	resumes-list/data-scientist-resume-example.pdf	KANDICE LOUDOR\n\nDATA SCIENTIST\n\nCONTACT\n\...	Github	(123) 456-7890	[kloudor@email.com]

Extract school

In [6]:
```
school_keywords = [
    'school',
    'college',
    'university',
    'academy',
    'faculty',
    'institute',
    'diploma',
]

def extract_education(input_text):
```

```
def extract_education(input_text):
    organizations = []

    for sent in nltk.sent_tokenize(input_text):
        for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(sen
            if hasattr(chunk, 'label'): #and chunk.label() == 'ORGANIZ
                organizations.append(' '.join(c[0] for c in chunk.leav

    education = set()
    for org in organizations:
        for word in school_keywords:
            if org.lower().find(word) >= 0:
                education.add(org)

    return education

df['school'] = df.text.apply(lambda x: extract_education(x))

df
```

Out[6]:

	path	text	name	phone	
0	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS	(123) 456-7890	[justin.green11@gr
1	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...	Stephen	(123) 456-7890	[stephen@beamj
2	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...	Ashley	(123) 456-7890	[ashley.do@gr
3	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...	Stephen	(123) 456-7890	[stephen@beamj
4	resumes-list/data-scientist-resume-example.pdf	KANDICE LOUDOR\n\nDATA SCIENTIST\n\nCONTACT\n\...	Github	(123) 456-7890	[kloudor@er
5	resumes-list/full-stack-developer-resume-examp...	ALEKS LUDKEE\nFull-Stack Developer\n\nludkee.a...	ALEKS	(123) 456-7890	[ludkee.aleks@er
6	resumes-list/shawon_resume.pdf	Mobile: +1 (709) 986-7643\nWebsite: https://sh...	Education	None	[sikamal@
7	resumes-list/entry-level-data-scientist-resume...	Trish Mathers\nEntry-Level Data Scientist\nInn...	Niantic Data Scientist Intern Seattle	(123) 456-7890	[tmathers@er

8	resumes-list/resume-example-option-college-stu...	Stephen\nGreet\nWeb Developer\n\nWork Experien...	Stephen	(123) 456-7890	[stephen@beamj
9	resumes-list/resume-example-option-nurse.pdf	ALICE LEWIS, APRN\n\nNurse Practitioner\n\nCON...	San Diego	(123) 456-7890	[alicelewis409@gr

Extract previous job titles

In [7]:
```
df_job_titles = pd.read_csv('job_titles_set.csv')
df_job_titles.title.values
```

Out[7]:
```
array(['owner', 'manager', 'president', ...,
       'corporate account executive', 'trade marketing',
       'library director'], dtype=object)
```

In [8]:
```
JOB_TITLE_DB = df_job_titles.title.values

def extract_job_titles(input_text):
    stop_words = set(nltk.corpus.stopwords.words('english'))
    word_tokens = nltk.tokenize.word_tokenize(input_text)

    #preprocessing
    filtered_tokens = [w for w in word_tokens if w not in stop_words]
    filtered_tokens = [w for w in word_tokens if w.isalpha()]

    grams = list(map(' '.join, nltk.everygrams(filtered_tokens, 2, 3)))

    found_skills = set()

    for i in filtered_tokens:
        if i.lower() in JOB_TITLE_DB:
            found_skills.add(i)

    for i in grams:
        if i.lower() in JOB_TITLE_DB:
            found_skills.add(i)

    return found_skills

df['job_titles'] = df.text.apply(lambda x: extract_job_titles(x))

df.head()
```

Out[8]:

	path	text	name	phone	email

0	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS	(123) 456-7890	[justin.green11@gmail.com]	
1	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...	Stephen	(123) 456-7890	[stephen@beamjobs.com]	{A
2	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...	Ashley	(123) 456-7890	[ashley.do@gmail.com]	
3	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...	Stephen	(123) 456-7890	[stephen@beamjobs.com]	{.
4	resumes-list/data-scientist-resume-example.pdf	KANDICE LOUDOR\n\nDATA SCIENTIST\n\nCONTACT\n\...	Github	(123) 456-7890	[kloudor@email.com]	

Part 2: Evaluation

Calculate similarity between job description and resume

In [9]:
```
job_description = open("job_description.txt", "r").read()
job_description
```

Out[9]:
```
"Software Developer\nLocation: St. John's;\n\nEach day our Software D
evelopers get to work on challenging problems. No two days are the sa
me, each day you'll collaborate with other Software Developers to pro
blem solve and write code that has an impact in the real world. Our p
roduct, Verafin, helps fight crime by stopping fraud and money launde
ring. Stopping the flow of this money means stopping crimes such as h
uman trafficking, elder abuse, and drug trafficking. Our Software Dev
elopers get the opportunity to move around the business as there are
new teams and projects developed all the time to help us towards our
mission of stopping crime. Being a Software Developer at Verafin mean
s getting the opportunity to have an impact on criminal activity by g
etting to do what you love – solve cool problems using code.\n\nEssen
tial Skills & Qualifications\nA university degree or college diploma
in Computer Engineering, Computer Science, or a combination of educat
ion and previous experience would be considered\nStrong analytical sk
ills for complex and creative problem solving\nExperience in object-o
riented software development       \nAutomated testing\nExcellent int
erpersonal and organizational skills; able to work closely with team
members\nWould be good to have experience in a few of the following a
reas\nJava\nExperience using JavaScript, CSS, REST\nPrevious experien
ce working with Core Banking Systems\nAmazon Web Services\nIntelligen
t systems, artificial intelligence and data science\nDistributed comp
uting\nDatabase technologies (PostgresSQL)\nBig data technologies\nDa
ta extraction, manipulation/cleansing and integration \n\n\nIndustry
and on-the-job training is provided for all roles at Verafin. \n\n\u2
00bVerafin places a high value on building a diverse team, candidates
of all backgrounds are encouraged to apply.\n\nMobile devices are not
supported for job applications currently. Please apply using a deskto
p device for the best user experience.\n\nPlease note: we frequently
see our jobs posted on job aggregators, which are essentially search
engines for jobs. Generally those sites ask you to use their sites to
apply for the posted job and they do not send us the application. As
a reminder, the the only way to apply for a job with Verafin is on ou
r site www.verafin.com/careers. We look forward to reviewing your app
lication."
```

In [10]:
```
new_row = pd.DataFrame({'path':'job_description', 'text': job_descript
df = pd.concat([new_row, df]).reset_index(drop = True)

df.head()
```

Out[10]:

	path	text	name	phone	email
0	job_description	Software Developer\nLocation: St. John's;\n\nE...	NaN	NaN	NaN
1	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS	(123) 456-7890	[justin.green11@gmail.com]
2	resumes-list/resume-example-option-project-man...	Stephen Greet\nProject Manager\nPMP certified p...	Stephen	(123) 456-7890	[stephen@beamjobs.com]
3	resumes-list/resume-example-option-attorney.pdf	Ashley Doyle, Esq\n\nashley.do@gmail.com\n\n(1...	Ashley	(123) 456-7890	[ashley.do@gmail.com]
4	resumes-list/resume-example-option-sales.pdf	Stephen Greet\nSales Associate\n\nWork Experie...	Stephen	(123) 456-7890	[stephen@beamjobs.com]

In [11]:
```
# Remove stop words and punctuations from text
stop_words_l=stopwords.words('english')

df['text_cleaned']=df.text.apply(lambda x: " ".join(re.sub(r'[^a-zA-Z]

tfidfvectoriser=TfidfVectorizer()
tfidfvectoriser.fit(df.text_cleaned)
tfidf_vectors=tfidfvectoriser.transform(df.text_cleaned)

similarities=np.dot(tfidf_vectors,tfidf_vectors.T).toarray()

for i in range(len(similarities[0])):
    df.loc[i, "similarity"] = similarities[0][i]

df.sort_values(by='similarity', ascending=False, inplace=True)

df = df.drop(0)
df.reset_index(drop=True, inplace=True)

df
```

Out[11]:

	path	text	name	phone	
0	resumes-list/full-stack-developer-resume-examp...	ALEKS LUDKEE\nFull-Stack Developer\n\nludkee.a...	ALEKS	(123) 456-7890	[ludkee.aleks@
1	resumes-list/shawon_resume.pdf	Mobile: +1 (709) 986-7643\nWebsite: https://sh...	Education	None	[sikam
2	resumes-list/resume-example-option-software-en...	justin.green11@gmail.com\n(123) 456-7890\nWash...	Github SKILLS	(123) 456-7890	[justin.green11@
	resumes-list/resume-	Stephen\nGreet\nWeb		(123)	

Ranking Output

In [12]:
```
df[['path', 'name', 'email', 'similarity']]
```

Out[12]:

	path	name	email	similarity
0	resumes-list/full-stack-developer-resume-examp...	ALEKS	[ludkee.aleks@email.com]	0.143581
1	resumes-list/shawon_resume.pdf	Education	[sikamal@mun.ca]	0.138904
2	resumes-list/resume-example-option-software-en...	Github SKILLS	[justin.green11@gmail.com]	0.101460
3	resumes-list/resume-example-option-college-stu...	Stephen	[stephen@beamjobs.com]	0.079581
4	resumes-list/resume-example-option-project-man...	Stephen	[stephen@beamjobs.com]	0.079037
5	resumes-list/data-scientist-resume-example.pdf	Github	[kloudor@email.com]	0.052557
6	resumes-list/entry-level-data-scientist-resume...	Niantic Data Scientist Intern Seattle	[tmathers@email.com]	0.050098
7	resumes-list/resume-example-option-nurse.pdf	San Diego	[alicelewis409@gmail.com]	0.030303
8	resumes-list/resume-example-option-sales.pdf	Stephen	[stephen@beamjobs.com]	0.028344
9	resumes-list/resume-example-option-attorney.pdf	Ashley	[ashley.do@gmail.com]	0.021063