تمرین پنجم

تحلیل لینک اخبار

شایان محمدی زاده سماکوش ۹۸۱۰۲۲۷۳

نونا قاضی زاده ۹۸۱۷۱۰۰۷

مقدمه

در این تمرین هدف پیاده سازس الگوریتم تحلیل لینک از جمله pagerank و HITS است در این تمرین دو سناریو در نظر گرفته شده است که پیاده‌سازی ما مبتنی بر سناریو دوم است در سناریو دوم عناوین اخبار را به عنوان گره در نظر می‌گیریم و در صورتی که دو عنوان خبر بیش یک مقدار مشخص دارای تعداد کلمات یکسان باشند به هم متصل می‌شوند. در نهایت با استفاده از الگوریتم تحلیل لینک مهم‌ترین اخبار ها را به عنوان خروجی می‌دهیم. لازم به ذکر است برای دقیق بودن جواب برای category های مختلف به صورت جداگانه انجام می‌دهیم بدین معنا که با گرفتن یک category مشخص مهم‌ترین خبرهای آن را به عنوان خروجی می‌دهیم.

پیاده‌سازی

پیش پردازش

در قسمت make dataset from dataframe دیتایی را که در بخش‌های قبلی کراول کرده بودیم را لود می‌کنیم و سپس روند پیش پردازش را انجام می‌دهیم ابتدا بخش title دیتا اخبارمان را نرمالایز می‌کنیم و سپس عملیات توکنایزیشن را انجام می‌دهیم بدین صورت که ابتدا عنوان خبر را به کلمات آن بخش بخش می‌کنیم سپس حروف اضافه را حذف می‌کنیم اما از آنجا که این stopword ها کامل نیستند و تمام کلمات اضافه و علائم نگارشی را ندارند بنابراین یک فایل دیگر ایجاد می‌کنیم و در این فایل سایر کلمات اضافه و علائم نگارشی که نیاز داریم را می‌افزاییم و بعد از مرحله توکنایزیشن این کلمات را حذف می‌کنیم و سپس با lemmatization و stemming کلمات به ریشه‌شان می‌بریم.

ساخت ماتریس شباهت

در این بخش ماتریس شباهت ساخته می‌شود، ساخت ماتریس شباهت به این صورت است که اگر n را معادل تعداد خبرها در نظر بگیریم ابتدا یک ماتریس n*n تمام صفر مقداردهی اولیه می‌کنیم. هر دو عنوان خبر با هم مقایسه می‌شوند بدین صورت که پس از پیش پردازش‌های ذکر شده تعداد کلمات مشترک میان دو خبر را محاسبه می‌کنیم و یک treshold در نظر می‌گیریم و اگر تعداد کلمات مشترک بین دو عنوان کمتر از treshold باشد در آن خانه ماتریس مقدار صفر و اگر این دو عنوان خبر یکسان بودند در آن خانه ماتریس صفر قرار می‌دهیم زیرا شباهت جمله با خودش زیاد است و به دست آوردن آن فایده‌ای ندارد (یعنی در خانه i,j ماتریس اگر i=j باشد مقدار صفر قرار می‌دهیم.) و اگر تعداد کلمات مشترک بیشتر از treshold باشد مجذور تعداد کلمات مشابه را به تعداد کل کلمات عنوان خبر تقسیم می‌کنیم بدین صورت ماتریس شباهت را می‌سازیم و یک وزن برای یال‌های وزن دار نسبت می‌دهیم.

```
news_num = len(selected_df)
words_set = [set(ls) for ls in selcted_removed_tokenized_words]
similarity_mat = np.zeros((news_num, news_num), dtype=float)
treshold = 4

for i in range(news_num):
    for j in range(news_num):
        intersect_len = len(words_set[i].intersection(words_set[j]))
        if intersect_len < treshold or i == j:
            similarity_mat[i][j] = 0
        else:
            similarity_matrix[i][j] = (intersect_len ** (1.2)) / len(words_set[i])
```

ساخت گراف

پس از ساخته شدن ماتریس مشابهت، سطرهای آن را normalize l1 می‌کنیم و با استفاده از networkx گراف مربوط به ماتریس مجاورت ساخته می شود

normalizing similarity matrix

```
similarity_mat_normalized = normalize(similarity_mat, norm='l1')
```

create graph

```
graph = nx.from_numpy_array(similarity_mat_normalized)
```

پیاده سازی الگوریتم page rank

در این قسمت بر روی گراف به دست آمده الگوریتم pagerank پیاده شده است.

```
page_rank = nx.pagerank(graph, alpha=0.9)
```

پیاده سازی الگوریتم HITS

در این قسمت الگوریتم HITS پیاده سازی شده است که hubs و authorities را خروجی می‌دهد.

```
hubs, authorities = nx.hits(graph)
```

تحلیل لینک با استفاده از tf-idf vectorizer

در این بخش به جای آنکه معیار شباهت تعداد کلمات مشابه باشد. بردار tf-idf عناوین خبر به دست آمده با هم شباهت گرفته می‌شوند. پس از ساخت ماتریس شباهت سطرهای آن را normalize l1 می‌کنیم و با استفاده از networkx گراف مربوط به ماتریس مجاورت ساخته می شود و الگوریتم page rank و HITS را روی آن اجرا می‌کنیم.

creating vocabulary and tfidf vectorizer

```
vocabulary = set()
for doc in selected_df.clean_text:
    vocabulary.update(doc.split(' '))
vocabulary = list(vocabulary)

vectorizer = TfidfVectorizer(ngram_range=(1,2), vocabulary=vocabulary, stop_words=None, norm='l2')
```

save vecotrizer and load

```
pickle.dump(vectorizer, open("./vectorizer.pickle", "wb"))
vectorizer = pickle.load(open("./vectorizer.pickle", 'rb'))
```

```
doc_term = vectorizer.fit_transform([' '.join(x) for x in selcted_removed_tokenized_words])
```

create tf-idf similarity matrix

```
tf_idf_similarity_mat = doc_term.dot(doc_term.T)
```

normalizing similarity matrix

```
tf_idf_similarity_mat_normalized = normalize(tf_idf_similarity_mat, norm='l1')
```

create tf-idf graph

```
tf_idf_graph = nx.from_numpy_matrix(tf_idf_similarity_mat_normalized.toarray())
```

implement page rank algorithm on tf-idf graph

```
tf_idf_page_rank = nx.pagerank(tf_idf_graph, alpha=0.9)
```

HITS algorihm on tf-idf

```
tf_idf_hubs, tf_idf_authorities = nx.hits(tf_idf_graph)
```

گرفتن خروجی مهم‌ترین خبرها

یک تابع get_top_n_news تعریف می‌کنیم که ایدی خبرهایی که مهم‌ترین هستند را خروجی می‌دهد

```
def get_top_n_news(values, n=5):
    top_n = np.argsort(list(values))[::-1][:n]
    return top_n
```

نمونه خروجی

به طور مثال اگر category خبر ما دانش باشد می‌دانیم که با توجه به پاندمی کرونا علم به سمت کشف واکسن رفته است. همچنین در خبرهای مربوط به دانش از آنجا که کلمه ایلان ماسک زیاد می آید بنابراین در گراف آن این نود با نودهای دیگر ارتباط بیشتری دارد و در نتیجه جز خبر مهم به شمار می رود:

Page rank

```
واکنش عجیب بایدن به اظهارات ایلان ماسک | توییت متقابل مدیرعامل تسلا همراه با یک صورت حساب!
--------------------------------
واکسن mRNA چینی اختصاصی کرونای امیکرون در امارات آزمایش می‌شود
--------------------------------
خبر خوش؛ پس از موج امیکرون پاندمی کرونا پایان خواهد یافت | کرونا به بیماری فصلی مانند آنفلوانزا تبدیل می‌شود
--------------------------------
شمار موارد جهانی عفونت آبله میمونی به حدود ۸۰۰ رسید| ویروس بیماری احتمالا درون آمریکا در حال انتشار است
--------------------------------
مشاوران سازمان غذا و داروی آمریکا قرص ضد کرونای شرکت مرک را توصیه می‌کنند
--------------------------------
```

HITS

Authorities

```
واکنش عجیب بایدن به اظهارات ایلان ماسک | توییت متقابل مدیرعامل تسلا همراه با یک صورت حساب!
--------------------------------
توییت نگران‌کننده ایلان ماسک درباره بحران در تسلا و اسپیس‌ایکس | پای اوکراین در میان است
--------------------------------
واکنش دلپذیر ایلان ماسک در توئیتر | فعلا همه چیز برای مدیرعامل تسلا خوب است
--------------------------------
توییت عجیب ایلان ماسک و هجوم کاربران به او
--------------------------------
توییت‌های عجیب ایلان ماسک ادامه دارد | این بار یک جمله مفهومی شاید همراه با دلداری!
--------------------------------
```

Hubs

توییت نگران‌کننده ایلان ماسک درباره بحران در تسلا و اسپیس‌ایکس | پای اوکراین در میان است

توییت عجیب ایلان ماسک و هجوم کاربران به او

توییت‌های عجیب ایلان ماسک ادامه دارد | این بار یک جمله مفهومی شاید همراه با دلداری!

واکنش دلپذیر ایلان ماسک در توئیتر | فعلا همه چیز برای مدیرعامل تسلا خوب است

واکنش عجیب بایدن به اظهارات ایلان ماسک | توییت متقابل مدیرعامل تسلا همراه با یک صورت حساب!

TF-IDF page rank

آبله میمونی احتمالا برای ماه‌ها یا سال‌ها انتشاری بی‌سروصدا در جهان داشته است

اشباع بازار جهانی واکسن کرونا| بزرگترین شرکت واکسن‌سازی جهان تولید واکسن آسترازنکا را متوقف می‌کند

چین کیت آزمایش آبله میمونی را آماده می‌کند | ساخت واکسن در طول یک سال ممکن است

روبات انسان‌نمای تسلا به بهره‌برداری می‌رسد | ماسک، روز هوش مصنوعی را تغییر داد

آمریکا شمار بیشتری واکسن آبله میمونی می‌خرد| شمار موارد بیماری در حال افزایش است

TF-IDF HITS

Authorities

اشباع بازار جهانی واکسن کرونا| بزرگترین شرکت واکسن‌سازی جهان تولید واکسن آسترازنکا را متوقف می‌کند

آبله میمونی احتمالا برای ماه‌ها یا سال‌ها انتشاری بی‌سروصدا در جهان داشته است

چین کیت آزمایش آبله میمونی را آماده می‌کند | ساخت واکسن در طول یک سال ممکن است

آمریکا شمار بیشتری واکسن آبله میمونی می‌خرد| شمار موارد بیماری در حال افزایش است

چین نخستین مرگ‌های از کرونا را پس از بیش از یک سال گزارش می‌کند

Hubs

اشباع بازار جهانی واکسن کرونا| بزرگترین شرکت واکسن‌سازی جهان تولید واکسن آسترازنکا را متوقف می‌کند

آبله میمونی احتمالا برای ماه‌ها یا سال‌ها انتشاری بی‌سروصدا در جهان داشته است

چین کیت آزمایش آبله میمونی را آماده می‌کند | ساخت واکسن در طول یک سال ممکن است

آمریکا شمار بیشتری واکسن آبله میمونی می‌خرد| شمار موارد بیماری در حال افزایش است

چین نخستین مرگ‌های از کرونا را پس از بیش از یک سال گزارش می‌کند

حال اگر category ما ورزش باشد می‌دانیم قطعا کلمات استقلال، پرسپولیس، فوتبال، قهرمانی، تیم ملی و ... در عناوین خبر زیاد می‌آید بنابراین در گراف آن نود آن ارتباط بیشتری با سایر نودها دارد.

Page rank

ادعای خبرساز بازیکن جنجالی عراق؛ به تیم ملی فوتبال ایران دعوت شدم! | همسرم ایرانی است و از لیگ این کشور باز هم پیشنها
د دارم

ببینید | شادی مهدی طارمی با پرچم ایران در جشن قهرمانی پورتو | ژست خاص ستاره ایرانی مقابل هواداران

ببینید | شوخی ترسناک لورکوزن با سردار آزمون | واکنش ستاره ایرانی به اقدام باشگاه آلمانی

رای دیدار جنجالی لیگ برتر اعلام شد | بازی پرسپولیس - تراکتور ۳ بر صفر و دو بازیکن محروم شدند

زمان عجیب اعلام رای بازی جنجالی لیگ برتر | ۲ بازیکن پرسپولیس به کمیته انضباطی دعوت شدند

HITS

Authorities

ادعای خبرساز بازیکن جنجالی عراق؛ به تیم ملی فوتبال ایران دعوت شدم! | همسرم ایرانی است و از لیگ این کشور باز هم پیشنها
د دارم

تصمیم جنجالی ملی‌پوش ایرانی برای رفتن به آمریکا | واکنش مدیر تیم‌های ملی به اقدام خبرساز

سوءقصد خبرساز در فوتبال ایران | واژگون شدن خودروی چهره جنجالی تصادفی نبود!

پشت پرده خروج دختر وزنه‌بردار ایران از هتل تیم ملی و استوری خبرساز مادرش | ورزشکار جوان فقط مدالش را برد!

بانوی ایرانی سرمربی تیم ملی عراق شد

Hubs

ادعای خبرساز بازیکن جنجالی عراق؛ به تیم ملی فوتبال ایران دعوت شدم! | همسرم ایرانی است و از لیگ این کشور باز هم پیشنها
د دارم

تصمیم جنجالی ملی‌پوش ایرانی برای رفتن به آمریکا | واکنش مدیر تیم‌های ملی به اقدام خبرساز

سوءقصد خبرساز در فوتبال ایران | واژگون شدن خودروی چهره جنجالی تصادفی نبود!

پشت پرده خروج دختر وزنه‌بردار ایران از هتل تیم ملی و استوری خبرساز مادرش | ورزشکار جوان فقط مدالش را برد!

بانوی ایرانی سرمربی تیم ملی عراق شد

TF_IDF page rank

مجیدی گرانترین مربی تاریخ فوتبال ایران شد | رقم نجومی قرارداد سرمربی سابق استقلال
---
۲ بازیکن استقلال و پرسپولیس به تیم ملی دعوت شدند

عکس | اولین قهرمانی بچه‌های فوتبال ایران در آسیا | تمجید AFC از تیم کاپیتان سابق پرسپولیس

هافبک پرسپولیس هم رفتنی شد | خداحافظی با سرخ‌ها بعد از بازگشت به ایران

عکس | استوری خبرساز ستاره سابق پرسپولیس با طعنه به گل‌محمدی | خوش‌آمدگویی به بازیکن جدید سرخ‌ها

TF_IDF HITS

Authorities

```
۲ بازیکن استقلال و پرسپولیس به تیم ملی دعوت شدند
--------------------------------
مجیدی گرانترین مربی تاریخ فوتبال ایران شد | رقم نجومی قرارداد سرمربی سابق استقلال
--------------------------------
بازیکن سابق تیم ملی و استقلال درگذشت
--------------------------------
هافبک پرسپولیس هم رفتنی شد | خداحافظی با سرخ‌ها بعد از بازگشت به ایران
--------------------------------
عکس | استوری خبرساز ستاره سابق پرسپولیس با طعنه به گل‌محمدی | خوش‌آمدگویی به بازیکن جدید سرخ‌ها
--------------------------------
```

Hubs

```
۲ بازیکن استقلال و پرسپولیس به تیم ملی دعوت شدند
--------------------------------
مجیدی گرانترین مربی تاریخ فوتبال ایران شد | رقم نجومی قرارداد سرمربی سابق استقلال
--------------------------------
بازیکن سابق تیم ملی و استقلال درگذشت
--------------------------------
هافبک پرسپولیس هم رفتنی شد | خداحافظی با سرخ‌ها بعد از بازگشت به ایران
--------------------------------
عکس | استوری خبرساز ستاره سابق پرسپولیس با طعنه به گل‌محمدی | خوش‌آمدگویی به بازیکن جدید سرخ‌ها
--------------------------------
```

به طور کلی شانس حضور جملات طولانی در میان top ها بیشتر است زیرا جمله مهم جمله‌ای است که اطلاعات بیشتری در آن باشد در جملات bottom معمولا جملات کوتاه که اطلاعات خاصی ندارند می‌آید

توجه: یک بررسی انجام داده‌ایم که ببینیم آیا ماتریس شباهت متقارن هست یا نه و طبق کد زیر متوجه می‌شویم که ماتریس شباهت متقارن نیست اما دلیل اینکه خروجی hub و authorities مخصوصا در جملات top یکسان است که ماتریس شباهت بر اساس تعداد اشتراک کلمات میان دو عنوان است و وقتی جمله i با j اشتراک زیادی داشته باشد منطقا j هم با i اشتراک زیادی خواهد داشت. می‌دانیم hub و authorities برای ارجاع تعریف می‌شود بدین صورت که بیانگر عناوینی که خیلی به آن ارجاع داده می‌شود یا زیاد به بقیه ارجاع داده است. این ارجاع اشتراک است و اشتراک هم دوسویه به همین دلیل است که hub و authorities تقریبا یکسان می‌شود.