



## گزارش پروژه پایانی – فاز اول

استخراج رابطه در زبان فارسی با بررسی وابستگی جهانی Seraji و PerDT

مبانی پردازش زبان و گفتار

استاد: دکتر بهروز مینایی بیدگلی

اعضای گروه:

دانیال بازمانده ۹۷۵۲۱۱۳۵

محمدحسین کریمیان ۹۷۵۲۱۴۶۸

## هدف پروژه:

در این پروژه هدف اصلی یافتن روابط معنایی در زبان فارسی است به این صورت که بین دو entity، یک رابطه استخراج می‌کنیم.

در این گام که گام اول پروژه است، ابتدا باید داده‌های زیادی به دست بیاوریم زیرا برای رسیدن به هدف نهایی، به مجموعه بزرگی از داده‌ها نیاز داریم. سپس باید این داده‌ها را normalize کنیم و با حذف stop words و پیدا کردن ریشه کلمات، تعداد تکرار شدن هر کلمه را بدست می‌آوریم و جمله‌ها را هم مشخص می‌کنیم.

برای به دست آوردن داده‌ها، باید Crawl کنیم که به معنای دنبال کردن لینک‌ها یا اصطلاحاً “فرزیدن” در وب سایت‌ها است. برای استخراج داده‌ها از اینترنت، از کتابخانه BeautifulSoup استفاده می‌کنیم که برای استخراج داده از فایل‌های html کاربرد دارد.

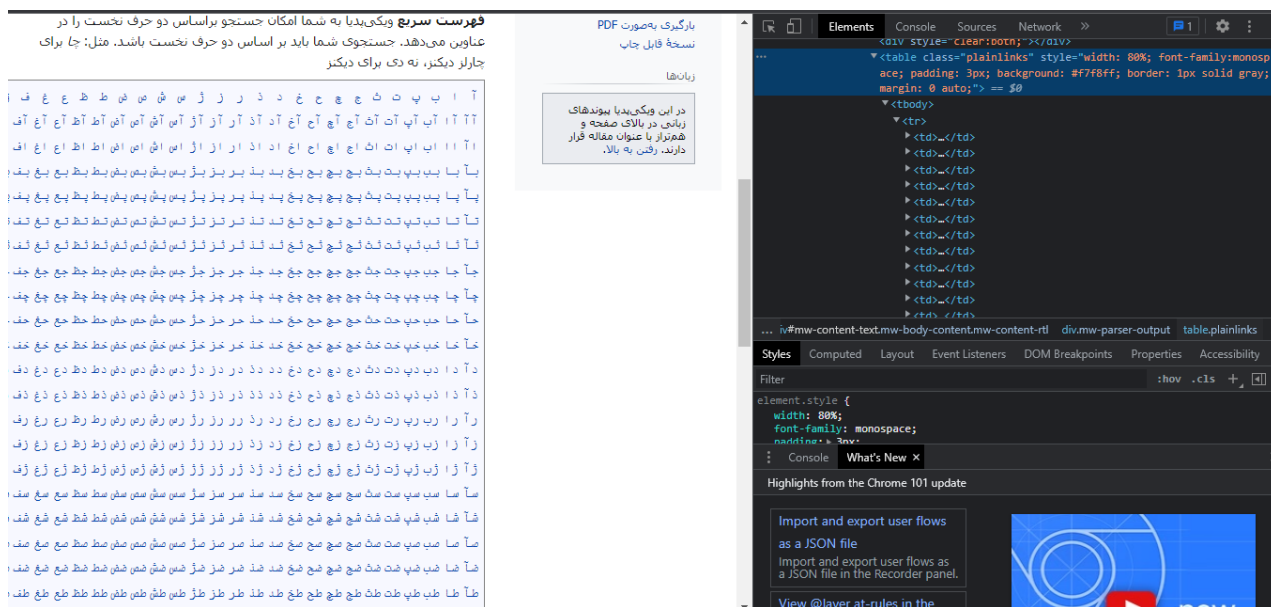
در این تمقیق از منابع سایت Wikipedia فارسی برای جمع‌آوری data استفاده کردیم به این صورت که ابتدا به فهرست منابع ویکی‌پدیا یک ریکوئست می‌فرستیم تا بتوانیم فهرست الفبای آن را پیدا کنیم و مطابق قطعه کد زیر، که در فایل collect\_alphabets.py می‌باشد، الفبای مورد نظر را که در ادامه قراره به لینک‌های مربوط به آن‌ها درخواست بزنیم، در یک فایل به نام alphabets.txt ذخیره می‌کنیم.

```
from bs4 import BeautifulSoup, Tag
import requests

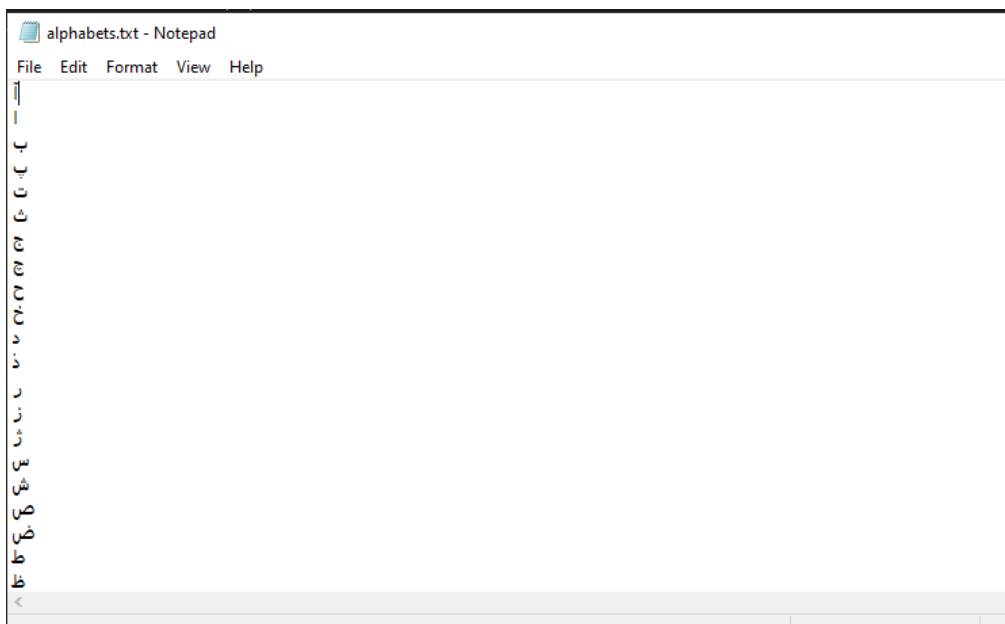
resultsFile = open("../data/raw/alphabets.txt", "w", encoding="utf8")
url = "https://fa.wikipedia.org/wiki/فهرست_سریع"
web = requests.get(url)

soup = BeautifulSoup(web.content, "html5lib")
table = soup.find('table', attrs={'style': 'width: 80%; font-family: monospace; padding: 3px; background: #f7f8ff;'})
for tr in table.contents[1].contents:
    if isinstance(tr, Tag):
        tds = tr.find_all("td")
        for td in tds:
            resultsFile.write(td.text)
```

مطابق شکل بالا وقتی به *url* مربوطه درخواست می دهیم، در صفحه *html* که وارد آن می شویم به دنبال جدولی می گردیم که دارای استایل مشخص شده باشد و همه موارد آن که مروف الفبای مورد نظر ما می باشند را در فایل alphabets.txt می نویسیم. استایلی که برای جدول مشخص کردیم با بررسی کردن تگ های این صفحه مانند شکل زیر به دست می آید.



حال در پوشه *data* که پوشه مربوط به داده هایمان است، در پوشه *raw* فایل *alphabets.txt* را چک می کنیم تا از اطلاعات ثبت شده مطمئن شویم.



مال که لیست مروفی از الفبا داریم، باید به مضمه هر مرف رفته و لینک های موضوعاتی که با آن مرف شروع می شوند را به دست می آوریم. به این خاطر که داده های مورد نیاز ما مداخلت سی هزار جمله می باشد، فقط موضوعاتی که در صفحه اول هر مرف هستند را در نظر می گیریم و همچنین مروفی از فهرست که بیش از یک مرف دارند را هم حذف می کنیم. سپس مانند قطعه کد زیر که در فایل *collect\_hrefs.py* می باشد، به این صفحات یک ریکوئست می فرستیم و آدرس صفحات مرتبط با آن ها را به دست می آوریم و در فایل *hrefs.txt* می نویسیم.

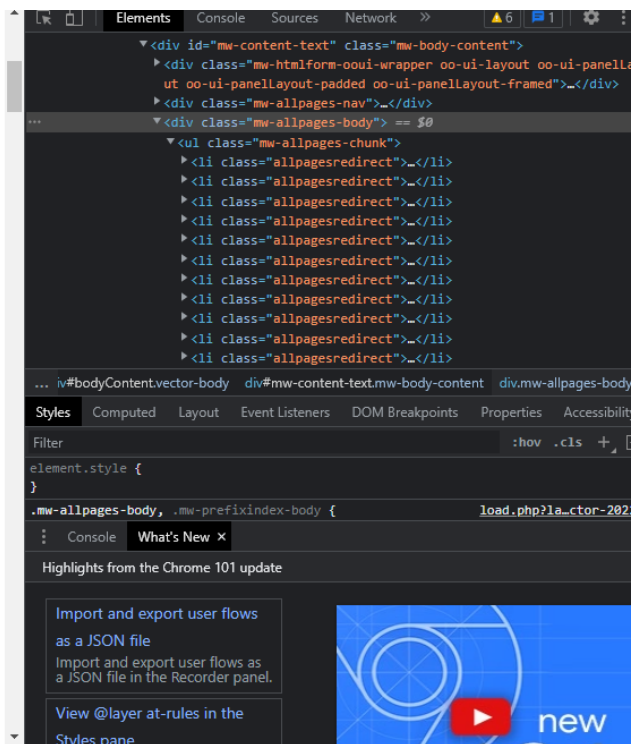
```
data_cleaner.py x collect_alphabets.py x script.py x collect_hrefs.py x _init_.py x
Visual layout of bidirectional text can depend on the base direction (View | Bidi Text Base Direction) Choose direction Hide notification Don't show again

7 url = "https://fa.wikipedia.org/wiki/ویژه:تمام_صفحه_ما"
8 alphabets = open('../data/raw/alphabets.txt', 'r', encoding="utf8")
9 hrefs = open('../data/raw/hrefs.txt', 'w', encoding="utf8")
10 count = 1
11
12 for alphabet in alphabets.read().splitlines():
13     if len(alphabet) == 1:
14         try:
15             # send a request to website
16             web = requests.get(url + 'from=' + alphabet + '&to=&namespace=0')
17         except:
18             # handle a problem that too requests to a website that gives us max_retries error
19             # This block code will get the url and retry 3 times if it gets error
20             session = requests.Session()
21             retry = Retry(connect=3, backoff_factor=0.5)
22             adapter = HTTPAdapter(max_retries=retry)
23             session.mount('http://', adapter)
24             session.mount('https://', adapter)
25
26             web = requests.get(url + 'from=' + alphabet + '&to=&namespace=0')
27
28             soup = BeautifulSoup(web.content, "html5lib")
29             table = soup.find('ul', attrs={'class': 'mw-allpages-chunk'})
30             for tr in table.contents:
```

مطابق کد بالا ابتدا فهرست الفبایی که از فایل آن می خوانیم را با *UTF8* اینکود می کنیم تا فرمت آن به دلیل فارسی بودن به هم نریزد. سپس به صفحه اول هر کدام از مروف الفبا در ویکیپدیا درخواست داده و با یافتن جدولی که کلاس مورد نظر ما را دارد و لیست موضوعات صفحه های ویکیپدیا در آن است، هر کدام از این موضوعات که در اسم آن مروفی مانند فاصله یا پرانتز است، یا درباره موضوع ابهام زدایی می باشد را نادیده گرفته و بقیه را *decode* می کنیم و در فایل *hrefs.txt* می نویسیم. کلاسی که برای جدول مشخص کرده ایم با بررسی آن صفحه به صورت زیر به دست آمده است.

صفحة قبلی (Zevtera) | صفحه بعد (اتومبلیستا)

آ.  
آ-10  
آ-10 تاندربولت  
آ-10 تاندربولت 2  
آ-4 اسكای هاوک  
آ-4 اسكای هاوک  
آ-4 اسكای هاوک  
آ-6 اینترودر  
آ-6 اینترودر  
آ-7 کرسر  
آ-7 کرسر  
آ-86929  
آ-سانتالول  
آ-لکس  
آ-لیست  
آ-لکس  
آ-لیست  
آ-می  
آ-می  
آ-نه-بی  
آ-ها  
آ-100  
آ-100 تاندربولت  
آ-100 تاندربولت 2  
آ-4 اسكای هاوک  
آ-4 اسكای هاوک  
آ-4 اسكای هاوک



مال برای اطمینان از نتایج، قسمتی از فایل *hrefs.txt* را در زیر می بینیم:



حال که لینک هایی که می خواهیم از آن ها اطلاعات بگیریم را ذخیره کردیم، باید به هر کدام جداگانه یک ریکوئست بزنیم و متنی که داخل آن نوشته شده است را استخراج کنیم و در پوشه *raw*، در فایل *raw\_dataset.json* ذخیره کنیم. برای این کار با *encode* کردن لینک های ذخیره شده در فایل *hrefs.txt* در مرحله قبل، به آن ها در فواست می دهیم و با کاوش کردن در فایل *html* سایت ها همه تگ های *p* آن ها که مربوط به متن است را به دست می آوریم. مطابق شکل زیر بعضی جمله ها که در همه مقاله ها می آید را نادیده می گیریم:

```
final_text = ''
for p_tag in div.findChildren("p", recursive=False):
    p_text = ''
    for element in p_tag.contents:
        if isinstance(element, Tag):
            if element.name != 'sup' and element.name != 'br':
                if 'می‌توانید با گسترش آن به ویکی‌پدیا کمک کنید' not in element.text \
                    and element.text.replace('\n', '') != '':
                    p_text += element.text
            else:
                if 'می‌توانید با گسترش آن به ویکی‌پدیا کمک کنید' not in element \
                    and element.replace('\n', '') != '':
                    p_text += element

    text = p_text.replace('\n', '')
    if p_text != '':
        final_text += p_text + ' '

dataset.append({'id': count, 'title': title, 'text': final_text})
count += 1
```

نتیجه آن را در فایل *raw\_dataset.json* مشاهده می کنیم:

ی فارسی و الفبای عربی و الفبای عبری و سایر ابجد‌های استفاده‌شده توسط زبان‌های سامی است. این حرف در زبان فارسی شروع کننده خیلی از اسم هاست : "text": "الف", "id": 1, "title": «جزو نشانه‌های سدوسه‌گانه‌ای که الفبای فارسی را تشکیل می‌دهند آورده شده‌است. تمام مطالب این بخش از لغت‌نامه ابن خلدون، مأخذ «هن»، شرح شده‌است؛ مگر آنکه از منبع دیگری نام برده شده باشد».

ابن کسب ۷ عنوان فهرستی در لیگ قهرمانی تیم ایالتیالی در اروپا و بعد از رنال مادرید دومین تیم پرفخار این رقابت‌ها شناخته میشود (ssotatf tsjo-ne k'altof m:lan) در سال ۱۹۰۸ میلادی در فرانسه به ریقای داخلی اینتر میلان و یوونتوس، در رتبه بالاتری قرار دارد .بن اتفاقات میلان تا فصل ۵۱-۱۹۵۰ موفق به کسب قهرمانی در ایالتیا نشد. در دهه ۵۰ میلان به اوج فوتبال ایالتیا با هدایت مشلت سوندی کرده-نولی-که متشکل از گوناړ کرن، گوناړ نورداړ و نیلس لیدیولم، بازگشت کرد. فرادادا یا باتولو مالینیو بود. در سال ۱۹۶۴، سیلیویو روسکونی، باشگاه را خریداری کرد و یک سال بعد با به خدمتگیری آریگیو ساکی به عنوان سرمربی، و بازیکنانی همچون مارکو فان باستن، رود گوئیوت و ف ن ان بدون شکست قهرمان ایالتیا کرد. میلان دو فصل بعد راه با قهرمان ایالتیا پشت سر گذاشت. کاپلو در اروپا هم افتدار ساکی را ادامه داد و همراه میلان، در ۳ فینال پایلیگ لیگ قهرمانان/جام باشگاهها حاضر شد پس، اما هر در ورزشگاه خانگی، بدترین نتیجه دوران بروسکونی است. میلان در اولین دهه قرن، یکی از قدرتمندترین باشگاههای اروپا و جهان بود. در فاصله سال‌های ۲۰۰۳ تا ۲۰۰۷ سه بار به فینال اروپا راه یافت شروع فصل ۰۷-۲۰۰۶ و جنم حضور این تیم در قهرمان اروپا در آن صادر کرد این جام محکم بود که از بازیگری، به کسر شصت امتیاز و حضور در مرحله بلق ایک لیگ قهرمانان اروپا، کاهش پیدا کرد. کاهش پیدا کرد. در سال ۲۰۰۷ یکی از صر اصلی در فینال بوکا جونویور ز با نتیجه ۲-۴ شکست داد و برایش نخستین بار این جام معتبر بین‌المللی را تحت نام جدید فتح کرد. میلان کفصل ۰۸-۲۰۰۷ را پایپینز از سطح انگظار به اتمام رسانده بود، در ایالتیا و اروپا بود. کانیز در پایان فصل با قراردادی به ارزش ۶۸٫۵ میلیون یورو به رنال مادرید پیوست. همچنین کارلو آچلوتی سرمربی میلان پس از ۸ فصل هدایت باشگاه و کسب ۱۱ جام، راهی جلیسی شد د بیرنه یوونتوس ترک گفت. کوچ پیرلو سرآغاز تغییر نسل طلایی بود که در طول ده سال، جاهایی معتبر داخلی و بین‌المللی را فتح کرده و میلان را به جایگاه پرافخرترین باشگاه در سطح بین‌المللی رسانده بودند به مه آنرا تغییر نام داد. در سال ۱۹۹۸ این ورزشگاه برای توسازی مدت کوتاهی تعطیل شد و پس از بازگشایی ظرفیت آن به حدود ۸۲ هزار نفر رسید. بسیاری از بازی‌های ایالتیا در همین ورزشگاه برگزار میشوند .این تنها بامی موفق به فتح جام شده‌است.۱۰۸ گروه‌های هواداری اوترای ایالتیا، در شهر میلان مستقر است. از این میان حمایت معتمد در حال حاضر بریکرت ریسورسره اصلیمترین گروه اوترای حامی میلان است ساب مداید. میلان در طول تاریخ خود از سال ۱۸۹۹ تاکنون، رؤسای گوناگونی داشته‌اند. برخی از این روسا هزمان مالک باشگاه هم بوده‌اند، درحالیکه تعدادی از آن‌ها ریاست افتخاری باشگاه را برعهده داشته‌اند. این تیوب طلای میلان در سال ۲۰۰۴ کسب کرد. چندتنی رباقی او کوکو و رونالدینیو بودند [bałs dɔɐ]. از گرفت: جانی ریورا، اولین بازیکن میلان و دومین پرافخرترین پرافخر تیب توپ طلای اورینتلظفر فرانسوی در طولانی‌ترین دوران مربیگری میلان در ۴۵۰ بازی، در اختیار اوست. سیلیویو بروسکونی، میلان را به بالاترین دور ریاست باشگاه و با ۲۱ سال ریاست در اختیار دارد. میلان پس از رنال تورالد و نیلس لیدیولم، پیشترین بار بین‌الملل باشگاه در برابر موندا بدست آمده‌است. میلان در یکی از دیدارهای فصل ۰۸-۱۹۱۲، موندا را با نتیجه ۰-۱ شکست داده‌است. بولونیا در فصل ۲۳-۱۹۲۲، سنگین‌ترین شکست را با ۸ گل به میلان تحویل کرده‌است ودو در سال ۱۹۹۶ با این شرکت کاملاً یکپارچه شد و تحت نام دایملر-بزرگ در آمدند. برخشی اهام یادماندهٔ آن‌ها به دایملر-کرسیلر-هوفاضا که از شرکت‌های وابسته دایملر-بزر بود و اخیرتر الحاق شدند (henau این زمان تولید شد، شروع سایتهٔ طولانی در امر تأمین تجهیزات الکتریکی حمل و نقل رهبال آلمان شد). لیه ایکوت به سازمانی این شرکت با محصولات و به اشتراک گذاری تبلیغات ویژه‌های طراحی مشترک گفت "title": 6, "id": ۰, "text": "۱۰). در دانشنامه ویکی‌پدیای انگلیسی، بازبینی‌شده در ۵ مارس ۲۰۱۵. «AEG». مشارکت‌کنندگان ویکی‌پدیا. ۱۱. سال ۱۹۹۴ به گذار آورد‌ه‌است AEG در طول عمر خانگی رسمه یکی از پرافخرترین تیمهای اروپاست و نسبت به ربقای داخلی اینتر میلان و یوونتوس، در رتبه بالاتری قرار دارد .۱۰). ایالتیالیی در اروپا و بعد از رنال مادرید دومین تیم پرفخار این رقابت‌ها شناخته میشود .بن اتفاقات میلان تا فصل ۵۱-۱۹۵۰ موفق به کسب قهرمانی در ایالتیا نشد. در دهه ۵۰ میلان به اوج فوتبال ایالتیا با هدایت مشلت سوندی کرده-نولی-که متشکل از گوناړ کرن، گوناړ نورداړ و نیلس لیدیولم، بازگشت کرد. فرادادا یا باتولو مالینیو بود. در سال ۱۹۶۴، سیلیویو روسکونی، باشگاه را خریداری کرد و یک سال بعد با به خدمتگیری آریگیو ساکی به عنوان سرمربی، و بازیکنانی همچون مارکو فان باستن، رود گوئیوت و ف ن ان بدون شکست قهرمان ایالتیا کرد. میلان دو فصل بعد راه با قهرمان ایالتیا پشت سر گذاشت. کاپلو در اروپا هم افتدار ساکی را ادامه داد و همراه میلان، در ۳ فینال پایلیگ لیگ قهرمانان/جام باشگاهها حاضر شد پس، اما هر در ورزشگاه خانگی، بدترین نتیجه دوران بروسکونی است. میلان در اولین دهه قرن، یکی از قدرتمندترین باشگاههای اروپا و جهان بود. در فاصله سال‌های ۲۰۰۳ تا ۲۰۰۷ سه بار به فینال اروپا راه یافت شروع فصل ۰۷-۲۰۰۶ و جنم حضور این تیم در قهرمان اروپا در آن صادر کرد این جام محکم بود که از بازیگری، به کسر شصت امتیاز و حضور در مرحله بلق ایک لیگ قهرمانان اروپا، کاهش پیدا کرد. کاهش پیدا کرد. در سال ۲۰۰۷ یکی از صر اصلی در فینال بوکا جونویور ز با نتیجه ۲-۴ شکست داد و برایش نخستین بار این جام معتبر بین‌المللی را تحت نام جدید فتح کرد. میلان کفصل ۰۸-۲۰۰۷ را پایپینز از سطح انگظار به اتمام رسانده بود، در ایالتیا و اروپا بود. کانیز در پایان فصل با قراردادی به ارزش ۶۸٫۵ میلیون یورو به رنال مادرید پیوست. همچنین کارلو آچلوتی سرمربی میلان پس از ۸ فصل هدایت باشگاه و کسب ۱۱ جام، راهی جلیسی شد د بیرنه یوونتوس ترک گفت. کوچ پیرلو سرآغاز تغییر نسل طلایی بود که در طول ده سال، جاهایی معتبر داخلی و بین‌المللی را فتح کرده و میلان را به جایگاه پرافخرترین باشگاه در سطح بین‌المللی رسانده بودند به مه آنرا تغییر نام داد. در سال ۱۹۹۸ این ورزشگاه برای توسازی مدت کوتاهی تعطیل شد و پس از بازگشایی ظرفیت آن به حدود ۸۲ هزار نفر رسید. بسیاری از بازی‌های ایالتیا در همین ورزشگاه برگزار میشوند .این تنها بامی موفق به فتح جام شده‌است.۱۰۸ گروه‌های هواداری اوترای ایالتیا، در شهر میلان مستقر است. از این میان حمایت معتمد در حال حاضر بریکرت ریسورسره اصلیمترین گروه اوترای حامی میلان است ساب مداید. میلان در طول تاریخ خود از سال ۱۸۹۹ تاکنون، رؤسای گوناگونی داشته‌اند. برخی از این روسا هزمان مالک باشگاه هم بوده‌اند، درحالیکه تعدادی از آن‌ها ریاست افتخاری باشگاه را برعهده داشته‌اند. این تیوب طلای میلان در سال ۲۰۰۴ کسب کرد. چندتنی رباقی او کوکو و رونالدینیو بودند [bałs dɔɐ]. از گرفت: جانی ریورا، اولین بازیکن میلان و دومین پرافخرترین پرافخر تیب توپ طلای اورینتلظفر فرانسوی در طولانی‌ترین دوران مربیگری میلان در ۴۵۰ بازی، در اختیار اوست. سیلیویو بروسکونی، میلان را به بالاترین دور ریاست باشگاه و با ۲۱ سال ریاست در اختیار دارد. میلان پس از رنال تورالد و نیلس لیدیولم، پیشترین بار بین‌الملل باشگاه در برابر موندا بدست آمده‌است. میلان در یکی از دیدارهای فصل ۰۸-۱۹۱۲، موندا را با نتیجه ۰-۱ شکست داده‌است. بولونیا در فصل ۲۳-۱۹۲۲، سنگین‌ترین شکست را با ۸ گل به میلان تحویل کرده‌است ودو در سال ۱۹۹۶ با این شرکت کاملاً یکپارچه شد و تحت نام دایملر-بزرگ در آمدند. برخشی اهام یادماندهٔ آن‌ها به دایملر-کرسیلر-هوفاضا که از شرکت‌های وابسته دایملر-بزر بود و اخیرتر الحاق شدند (henau این زمان تولید شد، شروع سایتهٔ طولانی در امر تأمین تجهیزات الکتریکی حمل و نقل رهبال آلمان شد). لیه ایکوت به سازمانی این شرکت با محصولات و به اشتراک گذاری تبلیغات ویژه‌های طراحی مشترک گفت "title": 6, "id": ۰, "text": "۱۰). در دانشنامه ویکی‌پدیای انگلیسی، بازبینی‌شده در ۵ مارس ۲۰۱۵. «AEG». مشارکت‌کنندگان ویکی‌پدیا. ۱۱. سال ۱۹۹۴ به گذار آورد‌ه‌است AEG در طول عمر خانگی رسمه یکی از پرافخرترین تیمهای اروپاست و نسبت به ربقای داخلی اینتر میلان و یوونتوس، در رتبه بالاتری قرار دارد .۱۰). ایالتیالیی در اروپا و بعد از رنال مادرید دومین تیم پرفخار این رقابت‌ها شناخته میشود .بن اتفاقات میلان تا فصل ۵۱-۱۹۵۰ موفق به کسب قهرمانی در ایالتیا نشد. در دهه ۵۰ میلان به اوج فوتبال ایالتیا با هدایت مشلت سوندی کرده-نولی-که متشکل از گوناړ کرن، گوناړ نورداړ و نیلس لیدیولم، بازگشت کرد. فرادادا یا باتولو مالینیو بود. در سال ۱۹۶۴، سیلیویو روسکونی، باشگاه را خریداری کرد و یک سال بعد با به خدمتگیری آریگیو ساکی به عنوان سرمربی، و بازیکنانی همچون مارکو فان باستن، رود گوئیوت و ف ن ان بدون شکست قهرمان ایالتیا کرد. میلان دو فصل بعد راه با قهرمان ایالتیا پشت سر گذاشت. کاپلو در اروپا هم افتدار ساکی را ادامه داد و همراه میلان، در ۳ فینال پایلیگ لیگ قهرمانان/جام باشگاهها حاضر شد پس، اما هر در ورزشگاه خانگی، بدترین نتیجه دوران بروسکونی است. میلان در اولین دهه قرن، یکی از قدرتمندترین باشگاههای اروپا و جهان بود. در فاصله سال‌های ۲۰۰۳ تا ۲۰۰۷ سه بار به فینال اروپا راه یافت شروع فصل ۰۷-۲۰۰۶ و جنم حضور این تیم در قهرمان اروپا در آن صادر کرد این جام محکم بود که از بازیگری، به کسر شصت امتیاز و حضور در مرحله بلق ایک لیگ قهرمانان اروپا، کاهش پیدا کرد. کاهش پیدا کرد. در سال ۲۰۰۷ یکی از صر اصلی در فینال بوکا جونویور ز با نتیجه ۲-۴ شکست داد و برایش نخستین بار این جام معتبر بین‌المللی را تحت نام جدید فتح کرد. میلان کفصل ۰۸-۲۰۰۷ را پایپینز از سطح انگظار به اتمام رسانده بود، در ایالتیا و اروپا بود. کانیز در پایان فصل با قراردادی به ارزش ۶۸٫۵ میلیون یورو به رنال مادرید پیوست. همچنین کارلو آچلوتی سرمربی میلان پس از ۸ فصل هدایت باشگاه و کسب ۱۱ جام، راهی جلیسی شد د بیرنه یوونتوس ترک گفت. کوچ پیرلو سرآغاز تغییر نسل طلایی بود که در طول ده سال، جاهایی معتبر داخلی و بین‌المللی را فتح کرده و میلان را به جایگاه پرافخرترین باشگاه در سطح بین‌المللی رسانده بودند به مه آنرا تغییر نام داد. در سال ۱۹۹۸ این ورزشگاه برای توسازی مدت کوتاهی تعطیل شد و پس از بازگشایی ظرفیت آن به حدود ۸۲ هزار نفر رسید. بسیاری از بازی‌های ایالتیا در همین ورزشگاه برگزار میشوند .این تنها بامی موفق به فتح جام شده‌است.۱۰۸ گروه‌های هواداری اوترای ایالتیا، در شهر میلان مستقر است. از این میان حمایت معتمد در حال حاضر بریکرت ریسورسره اصلیمترین گروه اوترای حامی میلان است ساب مداید. میلان در طول تاریخ خود از سال ۱۸۹۹ تاکنون، رؤسای گوناگونی داشته‌اند. برخی از این روسا هزمان مالک باشگاه هم بوده‌اند، درحالیکه تعدادی از آن‌ها ریاست افتخاری باشگاه را برعهده داشته‌اند. این تیوب طلای میلان در سال ۲۰۰۴ کسب کرد. چندتنی رباقی او کوکو و رونالدینیو بودند [bałs dɔɐ]. از گرفت: جانی ریورا، اولین بازیکن میلان و دومین پرافخرترین پرافخر تیب توپ طلای اورینتلظفر فرانسوی در طولانی‌ترین دوران مربیگری میلان در ۴۵۰ بازی، در اختیار اوست. سیلیویو بروسکونی، میلان را به بالاترین دور ریاست باشگاه و با ۲۱ سال ریاست در اختیار دارد. میلان پس از رنال تورالد و نیلس لیدی

```

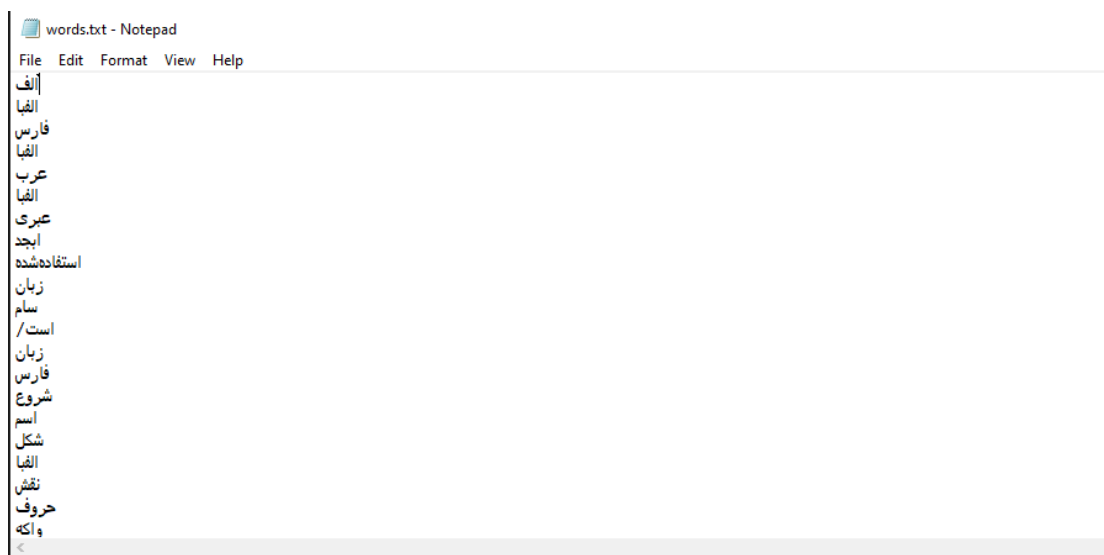
for row in raw_dataset:
    cleaned_text = normalizer.normalize(row.get('text'))
    cleaned_text_data.append({
        'id': row.get('id'),
        'text': cleaned_text,
        'subject': row.get('title')
    })

    for sentence in sent_tokenize(cleaned_text):
        sentences_data.append({
            'id': sentence_counter,
            'text': sentence,
            'subject': row.get('title')
        })
        sentence_counter += 1

    for word in word_tokenize(cleaned_text):
        new_word = lemmatizer.lemmatize(word)
        if new_word not in stop_words:
            if '#' in new_word:
                new_word = new_word.replace('#', '/')
            if new_word in counts:
                counts[new_word] += 1
            else:
                counts[new_word] = 1

```

سپس با *stemming* ریشه اصلی آن ها را به دست می آوریم و در نهایت چک می کنیم که این کلمات بین *stop words* های زبان فارسی هستند یا خیر و اگر در فایل *stop words* نبود، آن را در لیست کلمات در فایل *words.txt* می نویسیم که در شکل زیر این فایل را می بینیم.

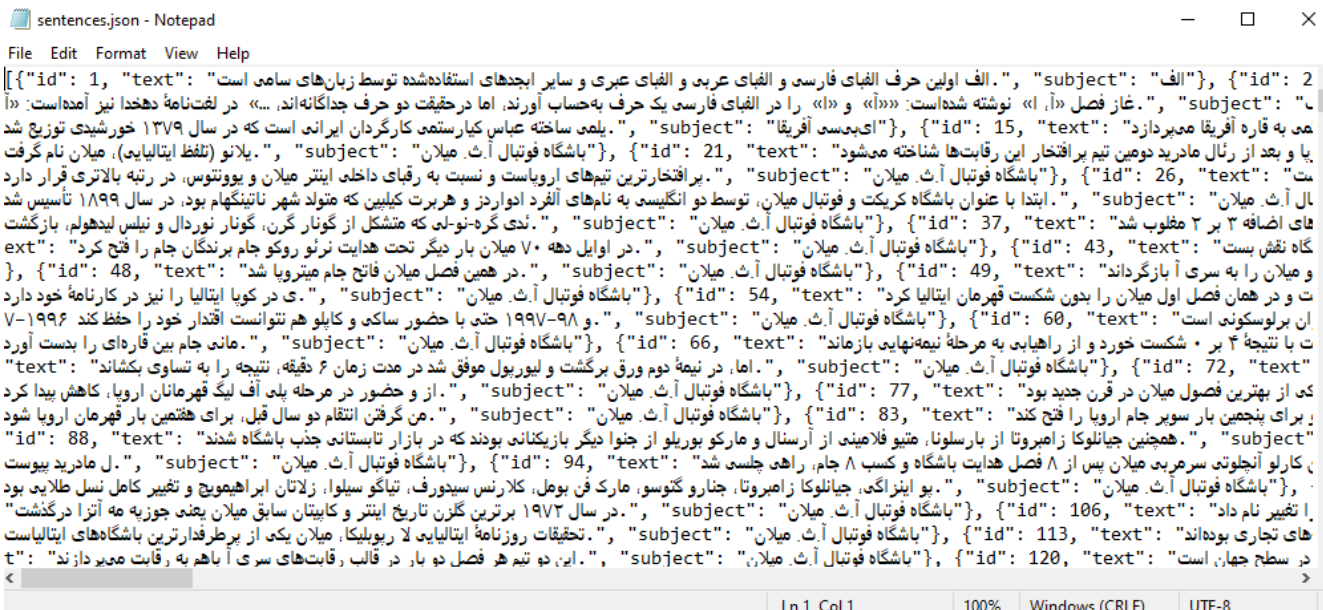




لیست *stop words* ها در پوشه *raw* در فایل *stopwords-fa.txt* قرار دارد. سپس هر جمله که توسط این

توابع جدا شده اند را در *sentences.json*، همراه با آیدی و موضوع آن می نویسیم. فرمت ذخیره شده را در

شکل زیر میں دیکھیں۔



همچنین برای این که هر کلمه را با تعداد تکرارش داشته باشیم، یک دیکشنری تعریف می‌کنیم و هر بار که کلمه

ای تکرار شد یکی به تعداد تکرار آن اضافه می کنیم و در نهایت آن را پس از *sort* کردن بر اساس تکرار بیشتر، در

فایل *count.txt* ذخیره می‌کنیم.

8162 : اشد/شو  
/8139 : است  
7173 : کرد/کن  
5650 : بود/باش  
4763 : سال  
3722 : داشت/دار  
2665 : ایران  
1914 : داد/ده  
1632 : شده/است  
1587 : گرفت/گیر  
1194 : شهر  
/1177 : هست  
1149 : کار  
1121 : کشور  
1086 : انگلیسی  
1052 : توانست/توان  
1042 : تاریخ  
1017 : گفت/گو  
932 : کتاب  
916 : بن  
882 : زمین  
807 : دست  
805 : یافت/باب  
766 : رفت/رو  
766 : آمد/آ  
741 : جهان  
734 : گروه  
722 : شرکت  
714 : نظر  
711 : آلمان  
710 : آب  
690 : تولید  
670 : می‌توان  
659 : دوران  
657 : گشت/گرد  
657 : علی

برای این که براساس تعداد مرتب کنیم، از قطعه کد زیر استفاده می کنیم:

```
words_count = open('../data/word_count/count.txt', 'w', encoding='utf8')

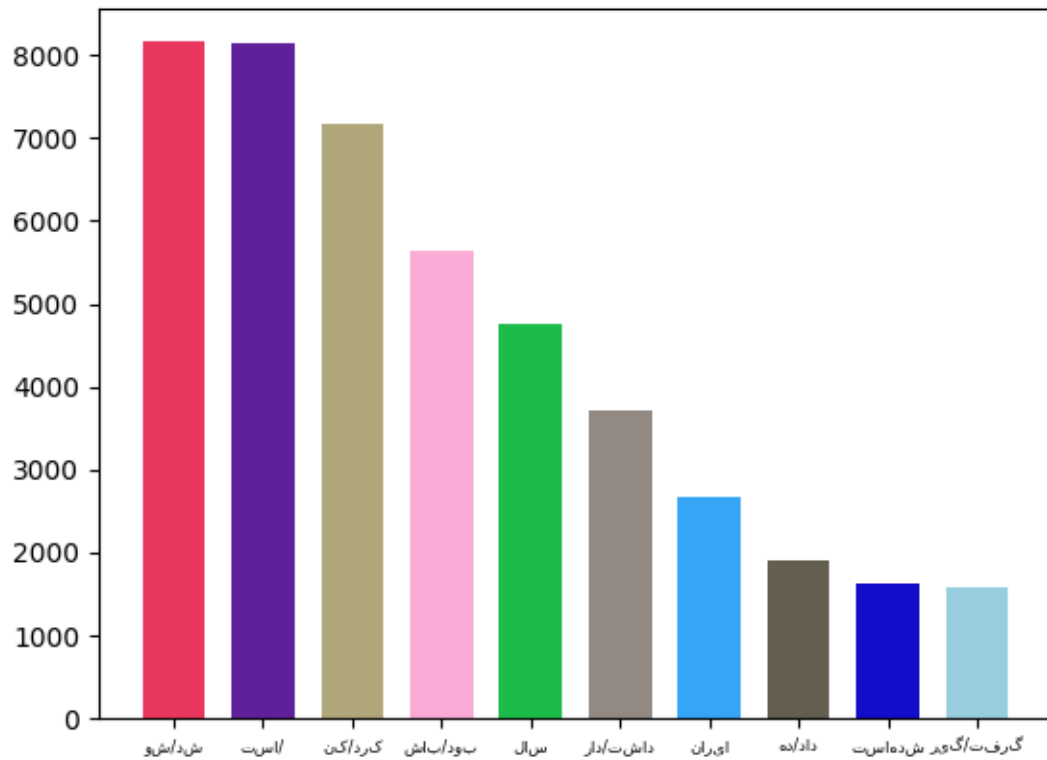
for i in sorted(counts, key=counts.get, reverse=True):
    words_count.write(i + " : "+str(counts[i])+"\n")

cleaned_dataset_file.write(json.dumps(cleaned_text_data, ensure_ascii=False))
sentence_broken_file.write(json.dumps(sentences_data, ensure_ascii=False))
```

حال که کلماتی که بیشترین تکرار را دارند پیدا کرده ایم، ده تا از آن ها که بیشترین تکرار را داشته اند در یک نمودار نمایش می دهیم. در فایل *plot.py*، ما با خواندن فایل *count.txt* و انتخاب ده مورد اول آن نمودار کلمات و تعداد تکرار آن ها را با رنگ های *random* رسم می کنیم و عکس نمودار را در پوشه *plot\_img* ذخیره می کنیم.

```
1 from __future__ import unicode_literals
2
3 import matplotlib.pyplot as plt
4 import numpy as np
5 import random
6 from bidi.algorithm import get_display
7 wordcount = open('../data/word_count/count.txt', 'r', encoding='utf8').read().splitlines()
8 x = []
9 y = []
10 for i in range(10):
11     row = wordcount[i].split(' : ')
12     y.append(int(row[1]))
13     x.append(get_display(row[0]))
14
15 rnd = []
16 for i in range(10):
17     r = random.random()
18     b = random.random()
19     g = random.random()
20     rnd.append((r, g, b))
21 plt.bar(x, y, width=0.7, bottom=None, align='center', data=None, color=_rnd_)
22 plt.xticks(fontsize=6)
23 |
24 plt.savefig('../data/plot_img/plot.png')
25
```

در نهایت در شکل زیر عکس نمودار را مشاهده می کنیم.



در انتها فایل *script.py*، همه توابع موجود برای گرفتن داده و نرمالایز کردن را انجام می دهد که کد آن در زیر نشان داده شده است.

```
data_cleaner.py × collect_alphabets.py × script.py × collect_articles.py × plot.py × collect_hrefs.py × __init__.py ×
1 import data_cleaner
2 import collect_alphabets
3 import collect_hrefs
4 import collect_articles
5
6 if __name__ == "__main__":
7     collect_alphabets.func()
8     collect_hrefs.func()
9     collect_articles.func()
10    data_cleaner.func()
11
```