

# به نام خدا

نام : حسین سیم چی

تمرین : شماره اول

درس : پردازش زبانهای طبیعی

در ابتدا با تولید یک متن ورودی فارسی شامل ۵۰ جمله که حدودا ۶۰۰ کلمه را دربردارد به تفاوتها و ویژگیهای عملکرد سه ابزار "استپ وان (nlp.sbu.ac.ir)" ، "پارسی پرداز" و "هضم" میپردازیم . سپس در انتها برای بیان ابزارهای اضافی و بیان کارهای بیشتر به بررسی کتابخانه ی NLTK و STANFORDNLP در محیط برنامه نویسی پایتون میپردازیم .

متن ورودی استفاده شده ( فارسی ) :

ستایش خداوندی را که در یگانگی، والا و در بی‌همتایی، نزدیک و در اقتدار شکوهمند ، و در ارکان خود بسی بزرگ است. دانشش بر همه چیز احاطه دارد و حال آنکه او در مقام خوش است و آفریدگان، همگی مقهور قدرت اویند. بزرگی که پیوسته بوده و ستوده‌ای که همیشه خواهد بود. خوب است آدمی جوری زندگی کند که آمدنش چیزی به این دنیا اضافه کند و رفتنش چیزی از آن کم... حضور آدمی باید وزنی در این دنیا داشته باشد باید که جای پایش در این دنیا بماند، آدم خوب است که آدم بماند و آدم تر از دنیا برود... نیامده ایم تا جمع کنیم آمده ایم تا عشق را ؛ ایمان را ؛ دوستی را ؛ با دیگران قسمت کنیم و غنی برویم... آمده ایم تا جای خالی را پر کنیم که فقط و فقط با وجود ما پر میشود و بس ! آمده ایم تا بازیگر خوب صحنه ی زندگی خود باشیم... پس بهترین بازی

خود را به نمایش بگذاریم . هر چقدر هم که گذشته تان آلوده بوده باشد، آینده تان هنوز حتی یک لکه هم ندارد. زندگی هر روزتان را با تکه شکسته های دیروزتان شروع نکنید!!! به عقب نگاه نکنید مگر اینکه چشم اندازی زیبا باشد. هر روز یک شروع تازه است. هر صبح که از خوابیدارمی شویم، اولین روز از باقی عمرمان است. یکی از بهترین راهها برای گذشتن از مشکلات گذشته این است.... که همه توجه و تمرکزتان را روی کاری جمع کنید که از خودتان در آینده برایش متشکر خواهید بود!!! باید راهی یافت، برای زندگی را زندگی کردن نه فقط زندگی را گذراندن.. باید راهی یافت، برای صبح ها با اُمید چشم گشودن، برای شب ها با آرامش خیال خوابیدن.. اینطور که نمیشود، نمیشود که زندگی را فقط گذراند ، نمیشود که تمام شدن فصلی و رسیدن فصلی جدید را فقط خُنکای ناگهانی هوا یادت بیاورد، نمیشود تا نوک دماغت یخ نکرده حواست به رسیدن پاییز نباشد.. اینطور پیش بروی یک آن چشم باز میکنی خودت را میان خزان زرد زندگی ات میابی ، و یادت هم نمی آید چطور گذرانده ای مسیر بهاری و سبز زندگی ات را.. اصلا خدا را هم خوش نمی آید ، راهت داده به دنیایش که نقشت را ایفا کنی، یک روز خوب حتی یک روز بد ، یک روز شیرین حتی یک روز تلخ ، یک روز آرام حتی یک روز پُرهیاهو ، وظیفه ی تو زندگی را با تمام و کمالش زندگی کردن است ، با تمام سِکانس های تلخ و شیرینش.. نمیشود که همه اش خسته باشی و سر سِکانس های تلخ بهانه بیاوری و گوشه ای به قهر کز کنی و بازی نکنی.. حق داری که خستگی ات را در کنی، اما حق نداری که دیگر مسیر را ادامه ندهی.. اینطور که نمیشود، تا دیر نشده باید راهی پیدا کرد، باید زندگی را زندگی کرد..

## توضیحاتی در رابطه با متن فوق :

متن استفاده شده از جمع اوری متون ادبی میباشد ، که در جمع اوری سعی شده است که از جملات پیچیده و حتی محاوره ایی جهت به چالش کشیدن ابزارها استفاده شود . از روی **عمد** در بعضی جاها از چندین علائم مانند ویرگول و نقطه استفاده شده است تا بهتر بتوانیم **نرمال سازی** را درک کنیم .

**استپ وان : ( \*تمامی تصاویر مراحل انجام کار به صورت جداگانه ارسال میگردد\* )**

زمانی که متن ورودی را وارد می کنیم در قسمت قطعه بندی برخی جملات را به **خوبی تفکیک** **نمیکند** ولی در بخش "**لغت-برچسب**" بسیار عالی عمل کرده و تمام بخش های جمله را به خوبی میشناسد .

در بخش مربوط به **استمر** ، چندین لغت را مثل "خداوندی" و "پرهیاو" به آن میدهیم و خروجی را که شامل **توضیحاتی در رابطه با ان اسامی** است مشاهده می کنیم .

**هضم : ( تمامی تصاویر مراحل انجام کار به صورت جداگانه ارسال میگردد )**

از جعبه ابزار هضم از طریق وب سایت [sobhe.ir/hazm](http://sobhe.ir/hazm) اقدام شده است . در بخش اصلاح متن یا تمیزکردن متن مشاهده می شود که **فاصله های اضافی ، ویرگول و نقطه ی اضافی را حذف میکند** و اصلاح شده ی آن را برمیگرداند .

در بخش **ریشه واژه ها** بسیار عالی عمل میکند و ریشه ی تمامی افعال را به درستی برمیگرداند در بخش **تحلیل صرفی** با تفکیک کردن تک تک اجزای جمله به خوبی هر قسمت و نوع آن را مشخص میکند .

در بخش **تجزیه ی نحوی** ، ابتدا تجزیه ی سطحی هر بخش یا جمله را مشخص میکند سپس تجزیه نحوی آنها را به شکل درخت و **ارتباط اجزای جمله با یکدیگر** مشخص میکند .

## بیان تفاوتها و ویژگی ها :

**دو ابزار فوق** هر کدام دارای ویژگی های بسیار عالی مخصوص خود هستند ولی در انتها به صورت کلی می توان گفت که جعبه ابزار **استپ وان** در بخش **برچسب گذاری برای هر لغت عالی عمل میکند** و دسته بندی خوبی را ارائه میدهد . جعبه ابزار **هضم** در بخش **تحلیل نحوی و صرفی** بسیار عالی عمل کرده و با بررسی درخت نحوی و بررسی ارتباط جملات فهم خوبی از جملات را انتقال میدهد .

### پارسی پرداز : (\*تمامی تصاویر مراحل انجام کار به صورت جداگانه ارسال میگردد\*)

در بخش **قطعه بندی** جملات بسیار عالی عمل میکند و مخصوصا در بخش word tokenization با تفکیک تک تک اجزا به خوبی جملات را شناسایی میکند . در بخش دیگر که به عنوان **ریشه یابی** در سایت میباشد ، زمانی که کلماتی مثل "خداوندی" و "پرهیاو" را به آن میدهیم به خوبی نوع آن را مشخص میکند و نکته بسیار جالب و کاربردی این است که **فرکانس رخداد هر کدام نیز برایمان مشخص میکند**.

در بخش دیگری نیز میتوان با ارائه ی فایل ورودی تک تک مراحل نامبرده شده را مشاهده کرد که این دسته بندی نیز بسیار مفید میباشد .

بخش **تجزیه معنایی** آن هنوز تمام نشده ولی در بخش **تجزیه نحوی** آن میتوان با ارائه جملات داخل متن تجزیه آن را بدست آورد

**نکته مهم** از تفاوتهای جعبه ابزار هضم و پارسی پرداز میتوان به این نکته اشاره کرد که بخش تجزیه نحوی جعبه ابزار هضم بسیار قوی تر کار میکند ،

زمانی که کل متن را به جعبه ابزار هضم میدهیم ساختار کلی تک تک جملات را رسم و مشخص میکند در حالی که در جعبه ابزار پارسی پرداز باید تک تک جملات را به آن بدهیم که کمی در افزایش زمان پردازش ، زمانی که متن ما طولانی باشد تاثیر گذار خواهد بود .

## کتابخانه ی STANFORDNLP در پایتون : ( سورس اصلی برنامه که شامل کد برنامه و عکس میباشد در پیوست ارسال میگردد )

یکی از کتابخانه های **چندزبانه** است که بیش از ۵۳ زبان را بررسی میکند . و بر اساس پیکره ی خانم دکتر سراجی نیز آموزش دیده شده است که ما نیز برای بررسی بخش Lemma و tokenization و تگ زدن بخشهای مختلف از آن استفاده میکنیم .

نصب و راه اندازی :

```
pip install stanfordnlp
```

نمونه کد برای پردازش زبان فارسی :

```
import stanfordnlp
```

```
LOCATION = r'C:\Data' :
```

محل ذخیره را مشخص میکنیم که در کامپیوتر من در فایل C و در پوشه Data ذخیره شده است

```
# stanfordnlp.download('fa', LOCATION, False) :
```

**این بخش را فقط برای بار اول دانلود می کنیم و نیازی به تکرار آن نیست ،**

با اجرای این خط از کد در پایتون پیکره ی خانم دکتر سراجی دانلود شده و از آن میتوانیم در بخش های دیگر استفاده کنیم

```
nlp = stanfordnlp.Pipeline(lang="fa", models_dir=MODELS_DIR,  
treebank='fa_seraji', use_gpu=False) :
```

بخش مربوط به تنظیمات میباشد که همانطور که بیان شد از زبان فارسی و پیکره خانم دکتر سراجی استفاده میکنیم .

```
doc = nlp(u"....."):
```

در داخل پرانتز هرگونه متن فارسی را جهت پردازش میتوان قرار داد

```
print(*[f'text: {word.text+" "}\tlemma: {word.lemma}\tupos: {word.upos}\txpos: {word.xpos}' for sent in doc.sentences for word in sent.words], sep='\n'):
```

این بخش مهم ترین بخش کد است و در آن Word tokenization و lemmatiation و تگ زنی هر

واژه انجام میشود

## جعبه ابزار NLTK در محیط JupyterLab :

در ابتدا باید فایل متنی که به فارسی در ابتدا ایجاد کردیم را در اینجا وارد کنیم سپس انرا Tokenize می کنیم همانند شکل زیر :

The screenshot shows the JupyterLab interface with a file browser on the left and a code editor on the right. The file browser shows a file named 'hossein.ipynb' modified 2 minutes ago. The code editor contains three code cells:

```
[1]: from nltk.tokenize import sent_tokenize
t = "hello my name is hossein simchi"
print(sent_tokenize(t))

['hello my name is hossein simchi']

[1]: from nltk.tokenize import word_tokenize
f = "سلام اسم من حسین است"
print(word_tokenize(f))

['سلام', 'و', 'اسم', 'و', 'من', 'و', 'حسین', 'و', 'است']

[4]: import nltk
file = open("C:\\Users\\Lenovo\\Desktop\\matn.txt", encoding = "UTF-8")
y = file.read()
print(y)
```

Below the code cells, there is a large block of Persian text:

ستایش خداوندی را که در یگانگی، والا و در بی-معتایی، نزدیک و در اقتدار شکوهمند، و در ارکان خود بسی بزرگ است. دانش پر همه چیز احاطه دارد و حال آنکه او در مقام خوش است و آفریدگان، همگی مقهور قدرت اویند. بزرگی که پیوسته بوده و ستوده-ای که همیشه خواهد بود. خوب است آدمی جوی زندگی کند که آمدنش چیزی به این دنیا اضافه کند و رفتنش چیزی از آن کم-حضور آدمی باید وزنی در این دنیا داشته باشد باید که جای پایش در این دنیا بماند، آدم خوب است که آدم بماند و آدم تر از دنیا برود نیامده ایم تا جمع کنیم آمده ایم تا عشق را ؛ ایمان را ؛ دوستی را ؛ با دیگران قسمت کنیم و غنی برویم -آمده ایم تا جای خالی را پر کنیم که فقط و فقط با وجود ما پر میشود و بس !آمده ایم تا بازبگر خوب صحنه ی زندگی خود باشیم هیچ بهترین بازی خود را به نمایش نگذاریم هر چند هم که گذشته-تان آلوده بوده باشد، آینده-تان هنوز حتی یک لکه هم ندارد. زندگی هر روزتان را با تکه شکسته های دیروزتان شروع نکنید!!!بیه عقب نگاه نکنید مگر اینکه چشم اندازی زیبا باشد. هر روز یک شروع تازه است. هر صبح که از خواب بیدار میشویم، اولین روز از باقی عمرمان است. یکی از بهترین راه ها برای گذشتن از مشکلات گذشته این است. که همه توجه و تمرکزتان را روی کاری جمع کنید که از خودتان در آینده برایش متفکر خواهید بود!!!باید راهی یافت، برای زندگی را زندگی کردن نه فقط زندگی را گذراندن ..باید راهی یافت، برای صبح ها با امید چشم گشودن، برای شب ها با آرامش خیال خوابیدن..اینطور که نمیشود، نمیشود که زندگی را فقط گذراند ، نمیشود که تمام شدن فصلی و رسیدن فصلی جدید را فقط خُکای ناگهانی هوا یادت بیاورد، نمیشود تا نوبت دعاغت بیغ نکرده حواست به رسیدن پاییز نباشد..اینطور بشو نباشی، یک آن خشم ساز میکنی خونت را میان خازن زود زندگی ات میریزی ،، یادت هم نبرد آید خطر

JupyterLab interface showing a file named `hossein.ipynb` open in the `Code` editor. The interface includes a file browser on the left, a toolbar at the top, and a status bar at the bottom.

The code in the cell [6] is:

```
[6]: from nltk.tokenize import sent_tokenize
print(sent_tokenize(y))
```

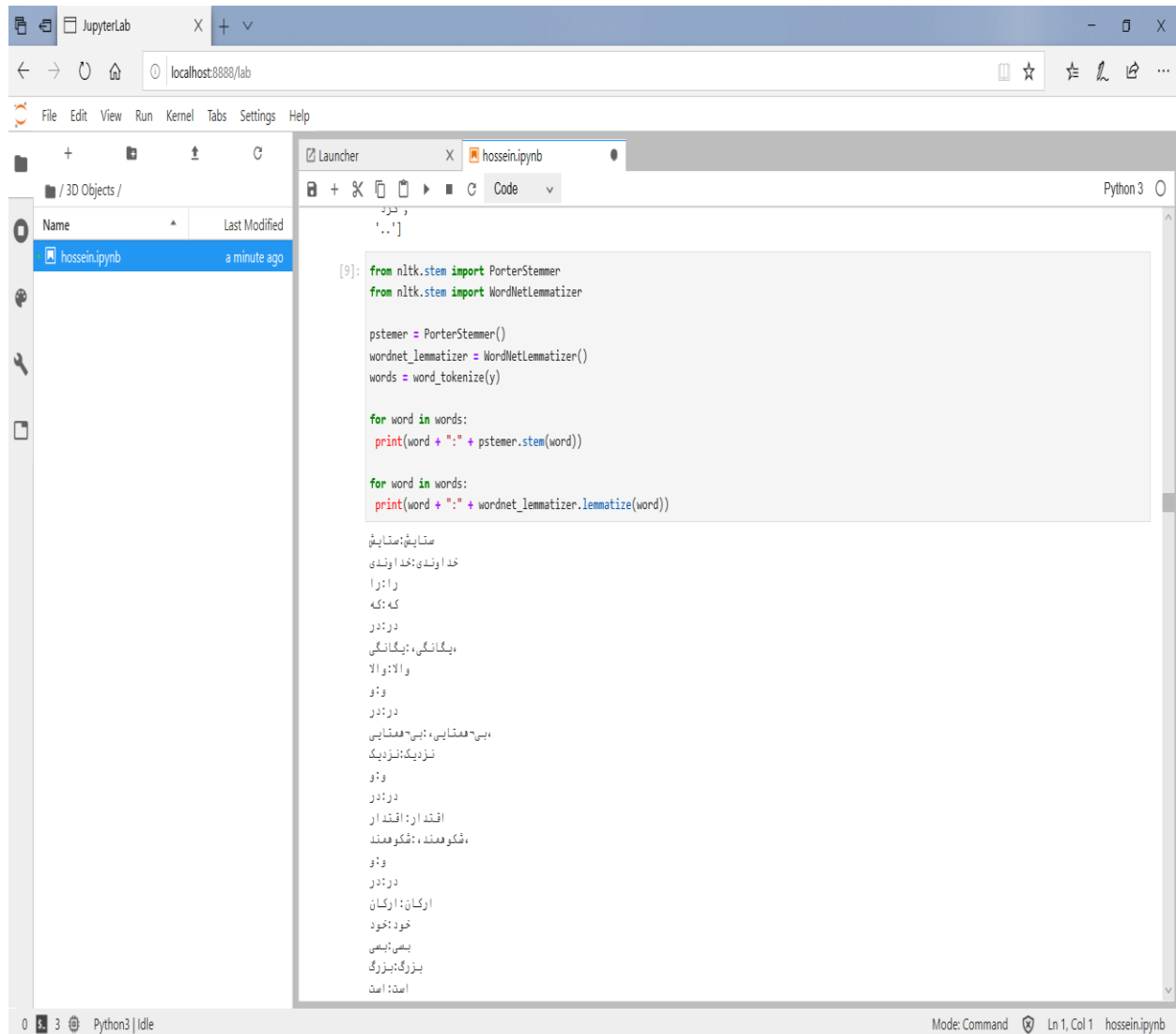
The output of the code is a list of sentences in Persian, displayed in cell [7]:

```
[7]: ['سنایش',
      'خداوندی',
      'و'را',
      'که',
      'در',
      'یگانگی',
      'والا',
      'و',
      'در',
      'بی-همنایی',
      'و نزدیک']
```

The status bar at the bottom indicates the file is `hossein.ipynb`, the mode is `Mode: Edit`, and the cursor is at `Ln 1, Col 1`.



و در ادامه به بررسی ریشه هر کلمه میپردازیم ، مطابق شکل زیر داریم :



The screenshot shows a JupyterLab environment with a file browser on the left, a code editor in the center, and a terminal at the bottom. The code in the notebook is as follows:

```
[9]: from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

pstemmer = PorterStemmer()
wordnet_lemmatizer = WordNetLemmatizer()
words = word_tokenize(y)

for word in words:
    print(word + ":" + pstemmer.stem(word))

for word in words:
    print(word + ":" + wordnet_lemmatizer.lemmatize(word))
```

The output of the code is a list of Persian words with their stems and lemmatized forms:

```
ستایش:ستایش
خداوندی:خداوندی
را:را
که:که
در:در
بیگانگی:بیگانگی
والا:والا
و:و
در:در
«بی»:«بی»
نزدیک:نزدیک
و:و
در:در
اقتدار:اقتدار
شکو:شکو
و:و
در:در
ارگان:ارگان
خود:خود
بسی:بسی
بزرگ:بزرگ
است:است
```

## نکات پایانی :

در این تمرین تمام تلاش خود را در جهت شناخت هرچه بیشتر ابزارهای متفاوت انجام داده ام ، در قسمت پیوست از تمامی مراحل انجام کار عکس گرفته شده و قرار داده شده است ، درضمن کد پایتون مربوط به دو کتابخانه ی بررسی شده نیز در فایل های متفاوت قرار داده شده است .  
باتشکر از زحمات شما ، حسین سیم چی

# پایان