گزارش تکلیف Cosine Similarity

درس داده‌کاوی

امیرحسین ابوالحسنی
۴۰۰۴۰۵۰۰۳

مقدمه

برای ارزیابی شباهت (Similarity) و یا تفاوت (Disimilarity) دو بردار از معیارهای متفاوتی می‌توان استفاده نمود. یکی از این معیارها تحت عنوان Cosine Similarity شناخته می‌شود که از این رابطه محاسبه می‌گردد:

$$\cos\theta = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}|| \cdot ||\vec{v}||}$$

که $\cos\theta$ همان Cosine Similarity می‌باشد.

حال با توجه به این متریک برای بررسی شباهت دو بردار، سعی در شباهت سنجی دو متن خبری فارسی می‌گردد. برای این کار سه خبر از خبرگزاری تسنیم تهیه شده است که موضوعات آنها طبق این جدول می‌باشد.

شماره خبر	نام خبر	موضوع
خبر ۱	آزمون: می‌دانیم چطور مقابل ازبکستان بازی کنیم	ورزشی - فوتبال - ایران
خبر ۲	جمع‌آوری ۳ "پتابایت" داده‌های ماهواره‌ای توسط سایت ورامین	فضا و نجوم
خبر ۳	بگیریستاین در پایان فصل منچسترسیتی را ترک می‌کند	ورزشی - فوتبال - جهان

در این تکلیف، هرکدام از این اخبار شباهتشان با یکدیگر سنجیده می‌شود. انتظار می‌رود دو خبر ورزشی بیشترین مقدار شباهت را داشته باشند.

کتابخانه‌ها

در این تکلیف برای بخش Web Scraping از کتابخانه‌های:

- Beautiful Soup
- Requests

و برای پردازش متن از کتابخانه‌های:

- hazm : پردازش متن فارسی
- nltk : شناسایی ترکیب‌های موردعلاقه
- re

و برای محاسبه Cosine Similarity از کتابخانه Numpy استفاده شده است.

Web Scraping

ابتدا با استفاده از کتابخانه Requests صفحه HTML گرفته شده، سپس با استفاده از کتابخانه Beautiful Soup این HTML پارس شده و متن خبر استخراج می‌گردد.

پیش پردازش متن

۱.۰ Normalization

ابتدا متن خام، نرمالایز [۱] می شود تا حروف و کلمات اضافی و غیرقابل استفاده از آن حذف شود.

۲.۰ Tokenization

در این مرحله، ابتدا متن را جمله جمله کرده، و سپس کلمه کلمه می‌کنیم تا تمامی کلمات موجود در متن را به شکل لیست در اختیار داشته باشیم.

۳.۰ برچسب گذاری دستوری

به هر کلمه یک برچسب تخصیص داده می‌شود که نشان دهنده نقش دستوری آن کلمه در آن جمله می‌باشد.
برای مثال:
اسم - **فعل** - **قید** - ...

بازیابی کلمات کلیدی

گرامرهایی تعریف می‌کنیم که عباراتی با این گرامر برای ما اهمیت دارند. برای مثال گرامر

NP: <NOUN.*><ADJ.*>? # Noun(s) + Adjective(optional)

به معنی عباراتی هستند که از یک اسم و یک صفت تشکیل شده اند.

Vectorization

به استفاده از مدل Sec2Vec دنباله کلمات کلیدی را به فضای برداری می‌بریم.

محاسبه شباهت کسینوسی

با توجه به فرمول ارائه شده و کتابخانه Numpy این مقدار را برای دو بردار محاسبه می‌کنیم.

نتیجه گیری

شباهت کسینوسی هر دو خبر به صورت زیر است:

	خبر ۱	خبر ۲	خبر ۳
خبر ۱	۱	۰.۳۴	۰.۳۶
خبر ۲		۱	۰.۲۲
خبر ۳			۱

همانطور که انتظار می‌رفت، خبر ۱ و ۳ بیشترین شباهت را نسبت به هر دو خبر دیگری دارند.

---

[۱] Normalize