مبانی پردازش زبان و گفتار

فاز اول پروژه
بررسی اشعار مولانا

نام تهیه‌کننده: هما سمسارها
استاد راهنما: دکتر صالح اعتمادی
تاریخ: ۱۴۰۰/۰۲/۲۰

فهرست مطالب

۱- منبع داده، ابزار، فرمت داده

داده از وبسایت گنجور به آدرس ganjoor.net جمع‌آوری شده‌است. به این‌صورت که اشعار مورد نیاز از url مربوطه crawl شده‌است. با بررسی فرمت url و نوشتن کد مناسب، از کتابخانه selenium پایتون برای دریافت متن استفاده شده‌است.

برای حفظ ساختار و اطلاعات اشعار، هر شعر در یک فایل به فرمت csv. ذخیره شده و عنوان شعر در ردیف اول قرار گرفته‌است. اشعار دفتر اول مثنوی در فولدر raw\masnavi-daftar-aval و اشعار دیوان شمس در raw\divan_shams قرار گرفته‌اند.

۲- پیش‌پردازش

- شکستن جملات: به‌دلیل قالب شعری داده، به‌جای جداسازی جملات مصراع‌ها را جدا کرده‌ایم. که در هنگام crawl فرمت ذخیره به‌گونه‌ای است که مصراع‌های اول و دوم هر بیت در ستون‌های مختلف قرار گرفته‌اند. درنتیجه فایل‌های موجود در فولدر raw به مصراع‌های خود شکسته شده‌اند و احتیاج به تغییر ندارند.
- شکستن کلمات: با شمستن جملات بر حسب کاراکتر space، کلمات جدا شده و کلمات هر مصراع در ستون‌های مختلف یک ردیف ذخیره‌شده‌اند. فرمت ذخیره csv. است، یعنی کلمات با کاراکتر , جدا شده‌اند.

۳- برچسب‌گذاری

واحد داده را بیت در نظر می‌گیریم. ابیات هر شعر در یک فایل قرار گرفته‌اند که نام فایل شماره شعر را مشخص می‌کند. داده از دو کتاب مثنوی و دیوان شمس جمع‌آوری شده و به این دو دسته تقسیم می‌شود. داده‌های در دسته در فولدرهای جدا قرار داده شده‌اند و نام هر فولدر برچسب را مشخص می‌کند.

۴- آمار

	مثنوی	دیوان شمس
تعداد واحد (بیت)	۴۰۱۳	۳۳۴۹
تعداد جمله (مصراع)	۸۰۲۶	۶۶۹۸
تعداد کلمه	۴۹۳۳۱	۴۹۸۲۱
تعداد کلمه منحصر به فرد	۸۹۱۱	۹۲۶۱
تعداد کلمه منحصر به فرد مشترک	۳۵۱۰	۳۵۱۰
تعداد کلمه منحصر به فرد غیرمشترک	۵۴۰۱	۵۷۵۱

- ۱۰ کلمه پر تکرار غیرمشترک هر برچسب:

مثنوی	ترا	گرچه	ازین	آنجا	ازو	هرچه	درین	فعل	اندرین	درمیان
تعداد تکرار	۶۳	۴۳	۴۰	۲۹	۲۴	۲۰	۱۹	۱۷	۱۷	۱۵
دیوان شمس	تویی	زهی	امشب	تبریز	بادا	جان‌ها	ساقیا	بی‌شما	جانا	کاین
تعداد تکرار	۷۷	۷۴	۶۷	۶۳	۵۸	۳۲	۲۷	۲۵	۲۳	۲۲

- ۱۰ کلمه مشترک هر برچسب بر حسب RNF:

مثنوی	ضد	خرگوش	جبر	قدرت	حس	فرمان	بخواهد	حکم	اثر	جهد
دیوان شمس	ساقی	دعا	مبادا	الصلا	لقا	بقا	صبا	خمار	ساغر	زنان

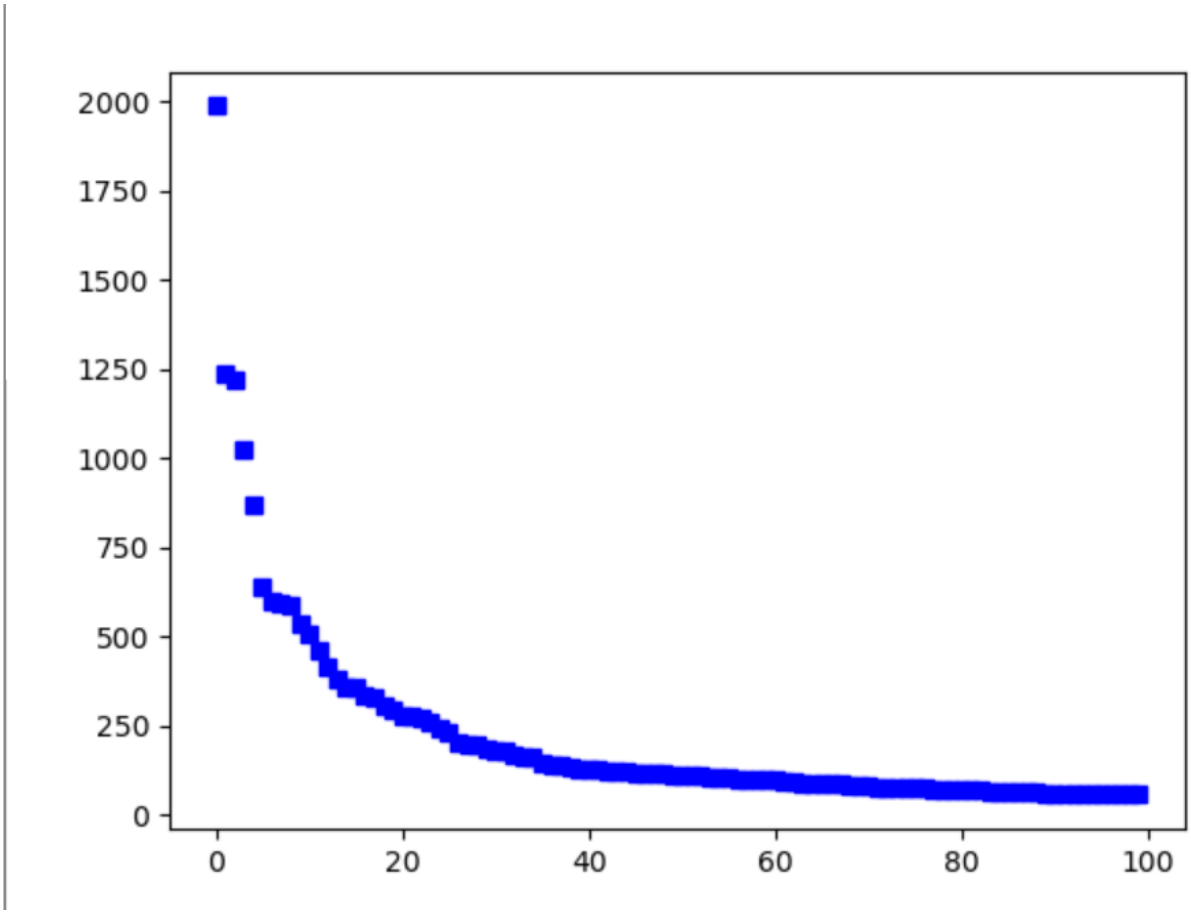
- ۱۰ کلمه برتر TF-IDF

مثنوی	گرچه	ازین	ازو	فعل	درمیان	ترا	اندرین	آنجا	هرچه
دیوان شمس	امشب	بادا	کاین	جانا	تویی	تبریز	ساقیا	بی‌شما	جان‌ها

- نمودار کلمات بر حسب تعداد تکرار:

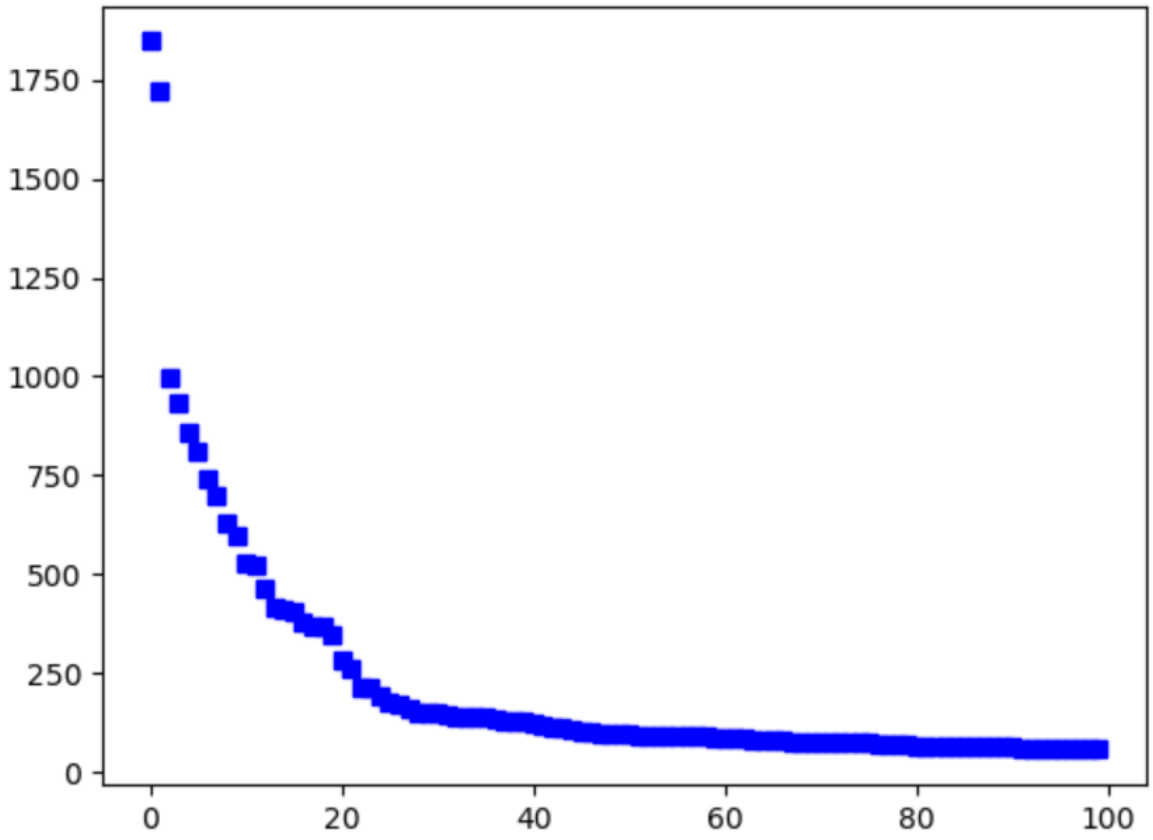
به‌دلیل تعداد زیاد کلمات و دیده نشدن آن‌ها در صورت نمایش همه، ۱۰۰ کلمه پر تکرار را نمایش داده‌ایم. به‌دلیل به‌هم ریختن متن فارسی در matplotlib، در نمودار فقط شماره کلمه را مشخص کردیم و خود کلمه را جداگانه نوشته‌ایم.

۱--**مثنوی:** و, را, از, در, آن, او, که, بر, چون, این, تو, تا, ز, من, ای, بود, هر, شد, گفت, با, ما, به, جان, خود, اندر, گر, نیست, حق, چه, پیش, هم, کرد, سر, آب, نه, باشد, دل, کو, آمد, سوی, دو, کند, پس, بی, شود, همچو, کی, یک, کن, چو, بهر, شیر, صد, دست, باز, جهان, زان, یا, آید, خویش, مر, نور, چونک, مرد, هست, بد,
جمله, زانک, آتش, پر, ور, دید, چشم, یکی, کان, رو, زین, کز, خدا, گشت, جز, بس, کار, مرا, دم, لیک, بعد, ترا, روی, نبود, وی, گوش, خلق, آنک, شه, وز, عقل, کین, روز, رنگ



شکل ۱: صد لغت پر تکرار مثنوی

۲- دیوان شمس: و, را, از, تو, که, در, ما, به, آن, ز, ای, این, جان, تا, بر, من, او, چو, چون, چه, دل, هر, شد, عشق, کن, مرا, با, سر, بود, خود, گر, هم, بیا, خوش, یا, آب, شود, صد, کند, همه, آمد, سوی, جا, چشم, اندر, جهان, گل, دست, دو, شمس, نه, چرا, شب, کجا, چنین, کی, روز, باشد, روی, پیش, اگر, یک, زان, کز, کو, فی, دم, عقل, گفت, تویی, مه, رو, خدا, مست, پر, زهی, نور, لا, نیست, جمله, گه, امشب, یار, می, باد, غم, وز, نی, هست, تن, تبریز, حق, مر, آتش, روح, پا, شاه, ساقی, زین, بادا



شکل ۲: صد لغت پر تکرار دیوان شمس