استخراج کلمات کلیدی

استاد درس

دکتر مینایی

کاری از

فربد صمصامی پور

احسان نژادیان

زمستان ۹۰

موضوع پروژه

هدف پروژه ارائه برنامه ایست که فایلی متنی را از ورودی دریافت کرده و کلمات کلیدی نمایانگر آن متن را استخراج کند.

ایده کلی پروژه

بعد از بررسی و تحقیق درباره روشهای استخراج کلمات کلیدی از متن، روش زیر را در نظر گرفتیم:

۱- در ابتدا کلمات متن ورودی جدا شده و stop word ها حذف میشوند. ما برای جدا سازی کلمات از کاراکتر فاصله (space) استفاده کردیم. روش مناسب تر برای این کار این است که بعد از جدا سازی کلمات، وجود ترکیبات آن کلمه با چند کلمه بعدی را در پایگاه داده ای از کلمات فارسی بررسی کنیم و طولانی ترین عبارت موجود را به عنوان کلمه متن می پذیریم و در صورتی که هیچ کدام از ترکیبات موجود نبود عبارت تک کلمه ای را در نظر میگیریم. برای مثال اگر کلمه "برنامه نویسی" در متن ورودی موجود باشد و کلمات را حداکثر دو بخشی فرض کنیم در این صورت برنامه عبارات "برنامه" و "برنامه نویسی" را مورد جستجو قرار میدهد و در صورتی که پایگاه داده کلمات بقدر کافی کامل باشد عبارت "برنامه نویسی" را انتخاب خواهد کرد. البته به دلیل نداشتن پایگاه داده مناسب به جدا سازی کلمات براساس کاراکتر فاصله اکتفا کردیم.

۲- پس از بدست آوردن واژگان متن آنها را براساس وابستگی معنایی خوشه بندی میکنیم تا در مواحل بعد از هر خوشه تعدادی کلمه را به عنوان کلمه کلیدی برگردانیم. هدف از خوشه بندی این است که کلمات کلیدی استخراج شده کل متن را از لحاظ معنایی پوشش دهند. معیار وابستگی معنایی کلمات را با توجه به مفهوم همرخدادی کلمات (coocurrence) و تعداد تکرار های کلمات (term frequency) در مجموعه ای از اسناد (به عنوان اسناد پشتیبان(corpus)) به صورت زیر در نظر گرفتیم:

$$\frac{coocurrence(A,B)_C}{termFrequency(A)_C + termFrequency(B)_C}$$

که دراین رابطه $coocurrence(A,B)_C$ همرخدادی کلمات A, B در مجموعه اسناد پشتیبان C و $termFrequency(A)_C$ تعداد تکرار کلمه A در مجموعه اسناد پشتیبان C و $termFrequency(B)_C$

تعداد تکرار کلمه B در مجموعه اسناد پشتیبان C است. بدیهی است که با افزایش تعداد معیار ها میتوان دقت خوشه بندی را افزایش داد.

با توجه به اینکه معیار وابستگی معنایی تعریف شده، صرفا فاصله دو کلمه را در فضای داده ها به دست میدهد تنها میتوان از روش hierarchical clustering برای خوشه بندی استفاده کرد. در مورد تعداد کلاستر ها نیز باید گفت به دلیلی مشابه، نمیتوان از معیار مرسوم ارزیابی خوشه بندی (مجموع مربعات فاصله داده ها از مرکز خوشه) استفاده کرد بنابراین ما به صورت تجربی تعداد خوشه ها را $\sqrt{n}$ در نظر گرفتیم که n تعداد واژگان متمایز سند ورودی است. همچنین ما از روش complete link برای خوشه بندی استفاده میکنیم زیرا به نظر میرسد، بهتر است وقتی دو کلاستر را ادغام کنیم که متفاوت ترین کلمات آنها از لحاظ معنایی، تا حد امکان شبیه باشند.

۳- پس از بدست آوردن خوشه ها، از هر خوشه تعدادی کلمه را که بیشترین مقدار tf×idf را در سند ورودی دارند برمیگردانیم. برای تعیین اهمیت خوشه ها از لحاظ معنایی، از میانگین tf×idf کلمات آن خوشه استفاده میکنیم. ما در این برنامه از خوشه هایی که این مقدار در آنها از میانگین اهمیتِ مهمترین خوشه و کم اهمیت ترین خوشه ، بیشتر است دو کلمه کلیدی و از سایر خوشه ها یک کلمه کلیدی انتخاب میکینم.

نحوه پیاده سازی

برای پیاده سازی این برنامه از زبان C++ به همراه Qt استفاده کردیم. همچنین فرض کردیم اسناد ورودی و اسناد پشتیبان به فرم utf-8 و فرمت txt. ذخیره شده اند. برنامه از کلاس های زیر تشکیل شده است:

۱- TextAnalyzer : این کلاس نام یک سند و بیشینه فاصله همرخدادی را به عنوان آرگومان های تابع سازنده اش میگیرد و همرخدادی ها و تعداد تکرارهای کلمات در سند را بدست می آورد. این کلاس در تحلیل متن stop word ها را حذف میکند که لیست stop word ها در فایلی در پوشه resource قرار دارد.

۲- CorpusData : این کلاس مقادیر برگشتی ار توابع کلاس TextAnalyzer را میگیرد و آن ها را ذخیره میکند. در واقع این کلاس اطلاعات بدست آمده از اسناد پشتیبان را نگهداری میکند. اطلاعات پس از بدست آمدن در پوشه ای به نام db که در کنار مسیر جاری پروژه ذخیره میشود. اطلاعاتی که در این کلاس ذخیره میشود به شرح زیر است:

جدولی از همرخدادی ها (فرض شده است در هر سطر کلمه اول از کلمه دوم کوچکتر است)، جدولی از تعداد تکرار کلمات، جدولی از تعداد اسناد شامل هر کلمه.

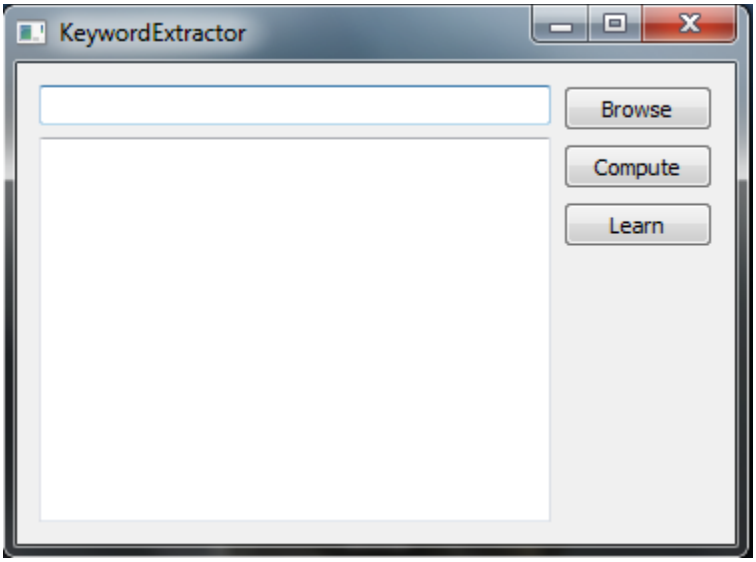
۳- ClusterMaker : این کلاس مجموعه ای از کلمات را در تابع سازنده اش دریافت میکند و با توجه به اطلاعات موجود در CorpusData آنها را با روش hierarchical clustering، با توجه به معیار وابستگی معنایی، خوشه بندی میکند.

۴- KeywordExtractor : این کلاس وظیفه انتخاب کلمات کلیدی از کلاسترهای بدست آمده را برعهده دارد.

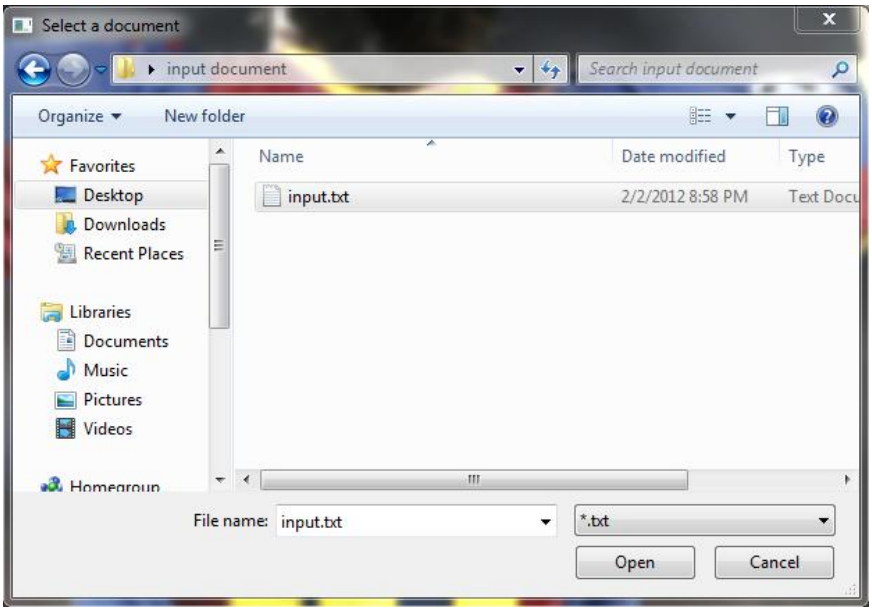
۵- Interface : این کلاس واسط گرافیکی برای کار با برنامه را ارائه میکند.

نحوه کار با برنامه

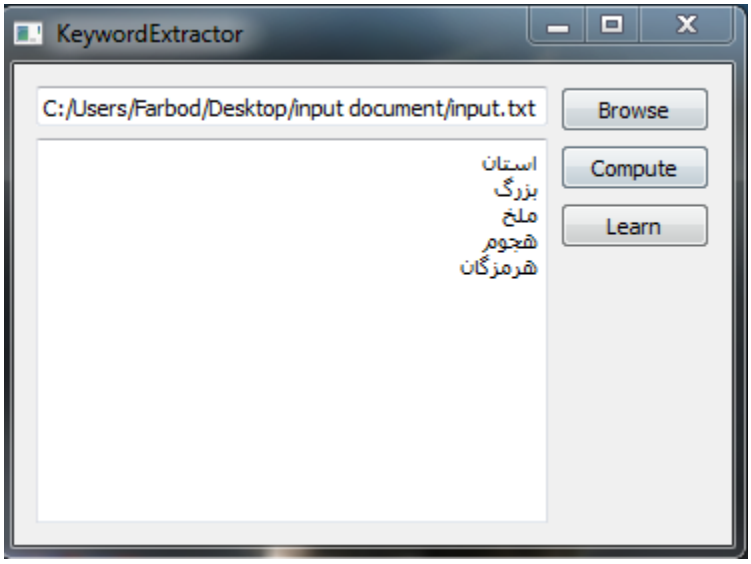
پس از اجرای برنامه با پنجره زیر مواجه میشویم:



در پنجره اصلی سه دکمه مشاهده میشود، دکمه Browse برای مشخص کردن سند ورودی میباشد که قرار است کلمات کلیدی آن استخراج شود. پس از فشردن این دکمه پنجره ای برای انتخاب فایل مورد نظر باز میشود.



پس انتخاب فایل ورودی با فشردن دکمه Compute برنامه شروع به استخراج کلمات کلیدی سند ورودی کرده و کلمات را در خروجی مطابق شکل نشان میدهد.



از دکمه Learn برای تغییر پایگاه داده استفاده میشود. با فشردن این دکمه میتوانید پوشه مربوط به اسناد دلخواه را انتخاب کرده و پس از تایید، برنامه اطلاعات مربوط به اسناد را با توجه به روش های توضیح داده شده ذخیره میکند. ما از قبل حدود ۵۴۰۰ سند از مجموعه اسناد همشهری را برای عمل یادگیری به برنامه داده ایم.

در زیر محتوای سند استفاده شده در مثال های بالا را مشاهده میکنید:

هجوم ملخ به هرمزگان

بندرعباس - خبرنگار همشهری: براثر هجوم دسته های بزرگ ملخ به استان

هرمزگان، خسارات زیادی به مزارع، باغها و نخیلات استان وارد شد.

هجوم ملخ بیشتر در مناطق قشم، کیش، هرمز، جاسک، میناب و رودان

مشاهده شده که بیش از ۲ هزار هکتار از فضای سبز را تهدید می کند.