# مستندات پروژه پنجم هوش محاسباتی

**اعضای گروه: الناز محمدی، زهرا رستمی**

## فاز اول: Preprocessing

**در این فاز پیش‌پردازش انجام می‌شود تا داده‌ها برای مراحل بعد آماده شوند. متن‌ها از HTML، اعداد، علائم نگارشی و کلمات بی‌معنی پاکسازی می‌شوند. همچنین کلمات به شکل پایه‌شان تبدیل می‌شوند. ستون‌های "Title"، "Body" و "Tags" پاکسازی می‌شوند. ردیف‌های خالی حذف می‌شوند و در نهایت داده‌های پاک‌سازی‌شده در فایل‌های جدید ذخیره می‌شوند.**

## فاز دوم: Word2Vec & Similarity Retrieval

**در این فاز یک مدل Word2Vec را آموزش می‌دهیم که پارامترهای آن به صورت زیر است:**

**– vector_size: بعد فضای بردار – window: اندازه پنجره متن**

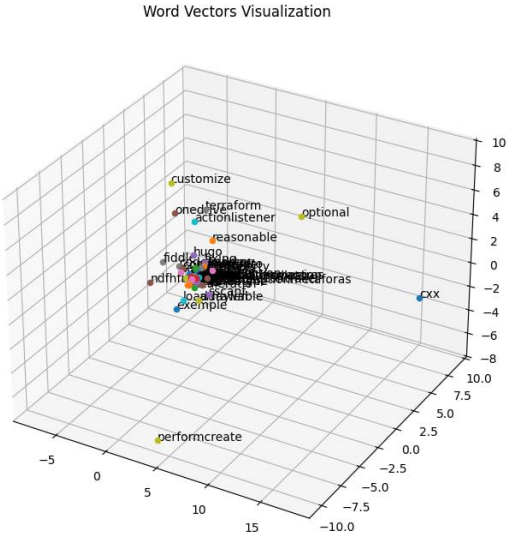
**– min_count: کلماتی که کمتر از این تعداد تکرار شده‌اند را نادیده می‌گیرد.**

**همچنین compute_sentence_vector بردار میانگین کلمات در یک جمله یا سند را با استفاده از مدل Word2Vec حساب می‌کند.**

**در ادامه برای نمایش بردارها ابعاد آن‌ها را با PCA به 3 بعد کاهش می‌دهیم.**

**نمودار بردارهای کلمات خوشه‌ها و روابط بین کلمات در فضای معنایی را نشان می‌دهد و نمودار بردارهای اسناد خوشه‌بندی اسناد بر اساس شباهت معنایی را نشان می‌دهد.**

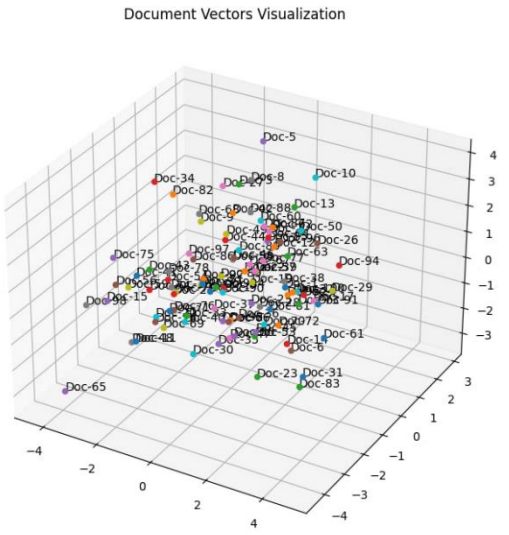
**خروجی بردار کلمات، برای کلمات نمونه‌گیری شده تصادفی به صورت زیر است:**




**– خوشه‌ها نشان می‌دهد کلمات معنایی مشابه دارند. مثلا، کلماتی مانند `customize`، `optional` و performcreate در فضاهای معنایی متمایزی قرار گرفته‌اند که نشان‌دهنده زمینه‌های منحصر به فرد آنهاست.**

**– بعضی از خوشه‌های مرکزی نشان می‌دهد همپوشانی معنایی زیاد در این کلمات وجود دارد.**

**همچنین خروجی بردار اسناد برای ۱۰۰ سند اول به صورت زیر است:**




**– اسناد خوشه‌های متراکمی تشکیل داده‌اند که نشان می‌دهد خیلی از سوالات از نظر معنایی به هم نزدیک هستند.**

**– نقاط دورافتاده مانند Doc-5 یا Doc-65 نشان می‌دهد موضوعات منحصر به فرد یا سوالاتی هستند که همپوشانی کمتری با بقیه دارند.**

**در ادامه یک کوئری به عنوان ورودی داده می‌شود و سپس 5 سوال مشابه با کوئری به همراه میزان شباهت آن‌ها محاسبه می‌شود.**

Enter your Question: SELECT datetime('now') is for current date.

Similar Questions:
1. setting time datetime (Similarity: 0.91)
2. get current datetime format h using node datetime library nodejs (Similarity: 0.84)
3. different class operated python like datetime added date (Similarity: 0.84)
4. converting string datetime datetime c (Similarity: 0.84)
5. get record count specified date range database record present one date return count c (Similarity: 0.84)

**خروجی نشان‌دهنده تطابق‌های مرتبط با پرس‌وجو است.**

**شباهت بالا برای این چند سوال (بالای 84 درصد) نشان می‌دهد که بسیاری از سوالات در مجموعه داده به جنبه‌های مشترکی از** datetime **می‌پردازند که می‌تواند برای بهبود نتایج جستجو مفید باشد.**

## فاز سوم: Tagging

در ابتدا 10٪ از مجموعه داده‌های validation را به طور تصادفی انتخاب می‌کنیم و فیلدهای متنی را تمیز و توکن‌بندی می‌کنیم همچنین بردارهای معنایی برای سوالات validation تولید می‌شود.

مدل Word2Vec روی مجموعه توکن‌بندی‌شده‌ی ترکیبی از title ها و body ها آموزش می‌بیند.

مدل KNN روی بردارهای معنایی داده‌های آموزشی آموزش می‌بیند و از فاصله کسینوسی به عنوان معیار استفاده می‌کند و 5 همسایه نزدیک را برای پیش‌بینی‌ها در نظر می‌گیرد.

در ادامه برای هر سوال validation، tag ها را با تحلیل tag های همسایه‌های نزدیک آن در مجموعه آموزشی پیش‌بینی می‌کند و Tag های همسایه‌ها را در یک لیست منحصر به فرد از tagهای پیش‌بینی‌شده ترکیب می‌کند.

در آخر بررسی می‌شود که چند سوال validation حداقل یک tag که به درستی پیش‌بینی شده دارند.

بر اساس خروجی کد: "Model Accuracy: 0.82" مدل به دقت 82٪ دست پیدا می‌کند که نشان می‌دهد حداقل یک tag را برای 82٪ از سوالات validation به درستی پیش‌بینی می‌کند.

همچنین پنج نمونه تصادفی ارائه می‌شود که اثربخشی سیستم پیش‌بینی tag را نشان می‌دهد:

```
Example 1:
Title: redux use one action type separate reducer
True Tags: redux, react-redux
Predicted Tags: reactjs, react-redux, ngrx, ngrx-store, javascript, react-router, angular, redux
Prediction Success: ✅
--------------------------------------------------
Example 2:
Title: enough storage space device store package starting android emulator
True Tags: android, xamarin, visual-studio-2015, mono
Predicted Tags: flutter, android-instant-run, android, android-studio, android-8.0-oreo, android-studio-2.1, visual-studio-code, android-emulator
Prediction Success: ✅
--------------------------------------------------
Example 3:
Title: initializing object assigning reference python
True Tags: python, oop
Predicted Tags: python-2.7, inheritance, class-attributes, attributes, python-3.x, class, methods, python, super, typeerror
Prediction Success: ✅
--------------------------------------------------
Example 4:
Title: enough storage space device store package starting android emulator
True Tags: android, xamarin, visual-studio-2015, mono
Predicted Tags: flutter, android-instant-run, android, android-studio, android-8.0-oreo, android-studio-2.1, visual-studio-code, android-emulator
Prediction Success: ✅
--------------------------------------------------
Example 5:
Title: e destructuring two object property name
True Tags: javascript, ecmascript-6, destructuring
Predicted Tags: function, php, c++, class, arrays, javascript, json, object, return, oop, organization
Prediction Success: ✅
--------------------------------------------------
```

مدل در پیش‌بینی حداقل یک tag واقعی در بیش‌تر موارد عملکرد خوبی دارد و روابط معنایی بین tagها را به طور موثر تشخیص می‌دهد، همان‌طور که در خوشه‌بندی فناوری‌های مرتبط مانند `redux`، `react-redux` و react-router مشاهده می‌شود.

همچنین می‌توان گفت برخی پیش‌بینی‌ها شامل tagهای نامرتبط (flutter, php) هستند که به دلیل هم‌پوشانی در فضای بردارهای معنایی رخ می‌دهد.مدل ممکن است بیش از حد تعمیم دهد و منجر به پیش‌بینی tagهای نامرتبط شود.

برای بهبود نتیجه می‌توان از فیلترهای اضافی یا وزن‌دهی برای اولویت‌دهی به tag ها که ارتباط قوی‌تری با همسایه‌های نزدیک دارند، استفاده کرد همچنین می‌توان اندازه مجموعه داده را افزایش داد یا نمونه‌های خاص‌تری در طول آموزش گنجانده شود تا نویز در پیش‌بینی‌ها کاهش یابد.