ورودی، متون خبری FardNews است.

ابتدا باید اخبار مورد نظر را دانلود و از حالت فشرده خارج کرد

هدف این است که با سه کتابخانه پایتون برای زبان فارسی Hazm,Parsivar,dadmatoolsعملیات نرمالسازی، توکن سازی در سطح کلمات و ریشه یابی را بر روی ورودی انجام دهیم.

**Hazm**

**نقاط قوت:**

به طور خاص برای فارسی طراحی شده است: Hazm به طور خاص برای پردازش متن فارسی (فارسی) طراحی شده است که آن را به ابزاری ارزشمند برای محققان و توسعه دهندگانی که با داده های زبان فارسی کار می کنند تبدیل می کند.

منبع باز و رایگان: Hazm یک کتابخانه منبع باز است که تحت مجوز MIT منتشر شده است، به این معنی که استفاده و اصلاح آن حتی برای مقاصد تجاری رایگان است.

قابلیت‌های مختلف NLP را ارائه می‌کند: Hazm طیف گسترده‌ای از قابلیت‌های NLP از جمله نشانه‌گذاری، ریشه‌گذاری، برچسب‌گذاری بخشی از گفتار، تجزیه و تحلیل احساسات را ارائه می‌کند که می‌تواند در زمان و تلاش برای توسعه‌دهندگان و محققانی که با متن فارسی کار می‌کنند، صرفه‌جویی کند.

پشتیبانی از رویکردهای مبتنی بر قانون و آماری: Hazm هم رویکردهای مبتنی بر قانون و هم رویکردهای آماری را برای برچسب گذاری و تجزیه بخشی از گفتار ارائه می دهد که می تواند در برنامه های مختلف NLP مفید باشد.

نصب و استفاده آسان: نصب و استفاده Hazm آسان است، و API آن به خوبی مستند شده است، و آن را هم برای مبتدیان و هم برای پزشکان با تجربه NLP در دسترس قرار می دهد.

**نقاط ضعف:**

پشتیبانی از زبان محدود: Hazm به طور خاص برای متن فارسی طراحی شده است، به این معنی که نمی توان از آن برای پردازش زبان های دیگر استفاده کرد.

دقت محدود برای برخی از وظایف NLP: در حالی که Hazm عملکردهای مختلف NLP را ارائه می دهد، دقت آن برای برخی از کارها مانند برچسب گذاری بخشی از گفتار و تجزیه ممکن است به اندازه سایر ابزارهای پیشرفته NLP نباشد.

پشتیبانی جامعه محدود: Hazm به اندازه برخی از کتابخانه های محبوب NLP به طور گسترده مورد استفاده یا پشتیبانی قرار نمی گیرد، به این معنی که یافتن کمک یا منابع برای عیب یابی مشکلات ممکن است دشوارتر باشد.

فعالیت توسعه محدود: به نظر می رسد که فعالیت توسعه Hazm در سال های اخیر کند شده است، به این معنی که ویژگی های جدید یا رفع اشکال ممکن است به اندازه برخی از کتابخانه های NLP دیگر رایج نباشد.

**Parsivar**

**نقاط قوت:**

به طور خاص برای فارسی طراحی شده است: پارسیوار به طور خاص برای پردازش متن فارسی (فارسی) طراحی شده است که آن را به ابزاری ارزشمند برای محققان و توسعه دهندگانی که با داده های زبان فارسی کار می کنند تبدیل می کند.

منبع باز و رایگان: پارسیوار یک کتابخانه متن باز است که تحت مجوز MIT منتشر شده است، به این معنی که استفاده و اصلاح آن حتی برای مقاصد تجاری رایگان است.

عملکردهای مختلف NLP را ارائه می دهد: پارسیور طیف گسترده ای از عملکردهای NLP از جمله نشانه گذاری، ریشه یابی، برچسب گذاری بخشی از گفتار، تجزیه و شناسایی موجودیت نامگذاری شده را ارائه می دهد که می تواند در زمان و تلاش توسعه دهندگان و محققانی که با متن فارسی کار می کنند، صرفه جویی کند.

از رویکردهای آماری استفاده می کند: پارسیور از رویکردهای آماری مانند مدل پنهان مارکوف (HMM) برای برچسب گذاری بخشی از گفتار استفاده می کند که ممکن است منجر به دقت بالاتر برای برخی از وظایف NLP در مقایسه با رویکردهای مبتنی بر قانون شود.

نصب و استفاده آسان: پارسیوار نصب و استفاده آسان است و API آن به خوبی مستند شده است و برای مبتدیان و متخصصان با تجربه NLP قابل دسترسی است.

**نقاط ضعف:**

پشتیبانی از زبان محدود: پارسیوار به طور خاص برای متن فارسی طراحی شده است، به این معنی که نمی توان از آن برای پردازش زبان های دیگر استفاده کرد.

دقت محدود برای برخی از وظایف NLP: در حالی که پارسیوار عملکردهای مختلف NLP را ارائه می دهد، دقت آن برای برخی از کارها مانند شناسایی موجودیت نامگذاری شده و تجزیه ممکن است به اندازه سایر ابزارهای پیشرفته NLP نباشد.

پشتیبانی محدود جامعه: پارسیوار به اندازه برخی از کتابخانه های محبوب NLP به طور گسترده مورد استفاده یا پشتیبانی قرار نمی گیرد، به این معنی که یافتن کمک یا منابع برای عیب یابی مشکلات ممکن است دشوارتر باشد.

فعالیت توسعه محدود: به نظر می رسد که فعالیت توسعه پارسیور در سال های اخیر کند شده است، به این معنی که ویژگی های جدید یا رفع اشکال ممکن است به اندازه برخی از کتابخانه های NLP دیگر تکرار نشود.

**Dadmatools**

**نقاط قوت:**

به طور خاص برای عربی طراحی شده است: DadmaTools به طور خاص برای پردازش متن عربی طراحی شده است، که آن را به ابزاری ارزشمند برای محققان و توسعه دهندگانی تبدیل می کند که با داده های زبان عربی کار می کنند.

منبع باز و رایگان: DadmaTools یک کتابخانه منبع باز است که تحت مجوز Apache 2.0 منتشر شده است، به این معنی که استفاده و اصلاح آن حتی برای مقاصد تجاری رایگان است.

عملکردهای مختلف NLP را ارائه می دهد: DadmaTools طیف گسترده ای از عملکردهای NLP را ارائه می دهد، از جمله نشانه گذاری، ریشه یابی، برچسب گذاری بخشی از گفتار، شناسایی موجودیت نامگذاری شده و تجزیه وابستگی، که می تواند در زمان و تلاش برای توسعه دهندگان و محققانی که با متن عربی کار می کنند، صرفه جویی کند.

از روش‌های یادگیری ماشینی استفاده می‌کند: DadmaTools از رویکردهای یادگیری ماشینی مانند فیلدهای تصادفی شرطی و ماشین‌های برداری پشتیبانی برای برخی از وظایف NLP استفاده می‌کند که ممکن است منجر به دقت بالاتر در مقایسه با رویکردهای مبتنی بر قانون شود.

نصب و استفاده آسان: نصب و استفاده از DadmaTools آسان است، و API آن به خوبی مستند شده است، و آن را برای مبتدیان و متخصصان با تجربه NLP در دسترس قرار می دهد.

**نقاط ضعف:**

پشتیبانی از زبان محدود: DadmaTools به طور خاص برای متن عربی طراحی شده است، به این معنی که نمی توان از آن برای پردازش زبان های دیگر استفاده کرد.

دقت محدود برای برخی از وظایف NLP: در حالی که DadmaTools عملکردهای مختلف NLP را ارائه می دهد، دقت آن برای برخی از وظایف مانند شناسایی موجودیت نامگذاری شده و تجزیه وابستگی ممکن است به اندازه سایر ابزارهای پیشرفته NLP نباشد.

پشتیبانی محدود از جامعه: DadmaTools به اندازه سایر کتابخانه های محبوب NLP به طور گسترده مورد استفاده یا پشتیبانی قرار نمی گیرد، به این معنی که یافتن کمک یا منابع برای عیب یابی مشکلات ممکن است دشوارتر باشد.

فعالیت توسعه محدود: به نظر می رسد که فعالیت توسعه DadmaTools در سال های اخیر کند شده است، به این معنی که ویژگی های جدید یا رفع اشکال ممکن است به اندازه برخی از کتابخانه های NLP دیگر رایج نباشد.