1. Инструментарий естественного языка (NLTK)

NLTK — это важная библиотека, поддерживающая такие задачи, как классификация, стемминг, маркировка, синтаксический анализ и семантическое рассуждение в Python. Это ваш основной инструмент для обработки естественного языка и машинного обучения. Сегодня он служит образовательной основой для разработчиков Python, которые только приступают к изучению NLP и машинного обучения.

Библиотека была разработана Стивеном Бердом и Эдвардом Лопером в Пенсильванском университете. Она сыграла ключевую роль в прорывных исследованиях NLP. NLTK, наряду с другими библиотеками и инструментами Python, теперь используют в своих учебных программах университеты по всему миру.

Библиотека довольно универсальна, однако (и это следует признать!) ее трудно использовать для обработки естественного языка. NLTK может быть довольно медленным и не соответствовать требованиям быстро развивающегося производственного использования. Кривая обучения очень крутая, но разработчики могут воспользоваться такими ресурсами, как эта полезная книга. Из нее вы узнаете больше о концепциях, лежащих в основе задач обработки языка, которые поддерживает этот инструментарий.

2. TextBlo

TextBlob является обязательным для разработчиков, которые начинают свое путешествие в NLP в Python. Идеально подходит для первого знакомства с NLP. TextBlob предоставляет новичкам простой интерфейс для помощи в освоении большинства основных задач NLP, таких как анализ настроений, POS-маркировка или извлечение именных фраз.

Мы считаем, что любой, кто хочет сделать свои первые шаги в направлении NLP с помощью Python, должен использовать эту библиотеку. Она очень полезна при проектировании прототипов. Однако она также унаследовала основные недостатки NLTK. Для эффективной помощи разработчикам, сталкивающимся с требованиями использования NLP Python в производстве, эта библиотека слишком медленная.

3. CoreNLP

Эта библиотека была разработана в Стэнфордском университете и написана на языке Java. Тем не менее, она оснащена оболочками для многих языков, включая Python, что делает ее полезной разработчикам, желающим попробовать свои силы в обработке естественного языка на Python. В чем заключается самое большое преимущество CoreNLP? Библиотека действительно быстра и хорошо работает в средах разработки продуктов. Кроме того, некоторые компоненты CoreNLP могут быть интегрированы с NLTK, что неизбежно повысит эффективность последнего.

4. Gensim

Gensim — это библиотека Python, которая специализируется на выявлении семантического сходства между двумя документами посредством векторного пространственного моделирования и инструментария тематического моделирования. Она может обрабатывать большие текстовые массивы с помощью эффективной потоковой передачи данных и инкрементных алгоритмов. Это больше, чем мы можем сказать о других пакетах, которые нацелены только на пакетную обработку и обработку в памяти. Что нам нравится в этой библиотеке, так это ее невероятная оптимизация использования памяти и скорость обработки. Все это достигается при помощи другой библиотеки Python, NumPy. Возможности векторного моделирования пространства этого инструмента также являются первоклассными.

5. spaCy

spaCy относительно молодая библиотека, предназначенная для производственного использования. Вот почему она гораздо доступнее других NLP-библиотек Python, таких как NLTK. spaCy предлагает самый быстрый синтаксический парсер, имеющийся сегодня на рынке. Кроме того, поскольку инструментарий написан на языке Cython, он также очень быстр и эффективен.

Но ни один инструмент не является совершенным. По сравнению с библиотеками, которые мы рассматривали до сих пор, spaCy поддерживает наименьшее количество языков (семь). Однако растущая популярность машинного обучения, NLP и spaCy как ключевой библиотеки означает, что этот инструмент может вскоре начать поддерживать больше языков программирования.

6. Polyglot

Следующая библиотека менее известна, но она относится к числу наших любимых библиотек, поскольку предлагает широкий спектр анализа и впечатляющий охват языка. Благодаря NumPy, она также работает очень быстро. Использование polyglot похоже на spaCy. Это очень эффективный, простой и в принципе отличный вариант для проектов, связанных с языками, не поддерживаемыми spaCy. Библиотека выделяется на фоне остальных еще и потому, что запрашивает использование выделенной команды в командной строке через конвейерные механизмы. Определенно стоит попробовать.

7. Scikit–learn

Эта NLP-библиотека удобна в использовании. Она предоставляет разработчикам широкий спектр алгоритмов для построения моделей машинного обучения. Ее функционал позволяет использовать метод «мешок слов» (bag-of-words model) для создания объектов, призванных решать задачи классификации текста. Сильной стороной этой библиотеки являются интуитивные методы классов. Кроме того, scikit-learn имеет отличную документацию, которая помогает разработчикам максимально использовать свои возможности.

Однако библиотека не использует нейронные сети для предварительной обработки текста. Поэтому, если вы хотите выполнять более сложные задачи предварительной обработки, такие как POS-маркировка для ваших текстовых корпусов, лучше использовать другие библиотеки NLP, а затем вернуться к scikit-learn для построения ваших моделей.

8. Pattern

Еще одна жемчужина среди библиотек NLP, используемых разработчиками Python для работы с естественными языками. Pattern предоставляет инструменты для частеречной разметки (part-of-speech tagging), анализа настроений, векторных пространств, моделирования (SVM), классификации, кластеризации, n-граммы поиска и WordNet. Вы можете воспользоваться преимуществами парсера DOM, веб-искателя, а также некоторыми полезными API, такими как API Twitter или Facebook. Тем не менее, этот инструмент по сути является веб-майнером и может оказаться недостаточным для выполнения других задач обработки естественного языка.