

1.1 Введение

Пакет **FreeLing** предоставляет функционал для анализа текста с учетом специфики языка.

FreeLing включает как библиотеку, так и исполняемый файл, позволяя пользователю анализировать текстовую информацию из командной строки.

Основные возможности FreeLing:

1. Разметка текста (токенизация)
2. Выделение предложений
3. Морфологический анализ
4. Определение составных слов
5. Вероятностное определение части речи неизвестного слова (hmm tagger)
6. Обнаружение и определение именной группы
7. Классификация именной группы
8. Построение дерева зависимостей (слов в предложении)
9. Определение местоимений (местоименных словоформ)
10. Нормализация и определение дат, чисел, процентных соотношений, валюты и физических величин (скорость, вес, температура, плотность и т.д.)
11. Определение части речи (вероятностное)

В настоящее время поддерживаемые языки: испанский, каталонский, галисийский, итальянский, английский, валлийский, португальский, австрийский, русский.

1.2 Лицензия

FreeLing распространяется под GNU General Public License (GPL).

Если предполагается встраивать пакет в ПО для дальнейшей установки конечному пользователю, необходимо согласовать условия, используя контактную информацию на сайте.

1.3 Вклад

Основы проекта заложены в исследовательском центре Каталонского политехнического университета (<http://talp.upc.edu>). В дальнейшем развитии участвовало множество людей, подробную информацию можно найти по адресу: (<http://www.lsi.upc.edu/~nlp/freeling>)

2. Инсталляция

Со всеми подробностями инсталляции можно ознакомиться на страницах англоязычной версии документации. Так же существует экспериментальная сборка под msvc 10.0, описание к которой находится в Readme, в соответствующей папке проекта.

3. Описание основного функционала.

1. Токенайзер (Tokenizer)

Используется для преобразования текста в набор токенов (слов, сокращений ...)

Для задания соответствующих правил используется файл настройки, содержащий регулярные выражения и список слов сокращений.

2. Сплиттер (Splitter)

Сплиттер на вход получает список токенов (возможно полученный на предыдущем шаге) и возвращает список предложений.

3. Модуль определения чисел.

На вход принимает предложение (после сплиттера), на выходе аннотированное, в соответствии с правилами, предложение. Для определения используются конечные автоматы.

4. Модуль определения даты.

Тоже самое, что и для чисел. На входе предложение — на выходе аннотированное измененное предложение с нормализованными датами.

5. Модуль поиска по словарю.

Модуль поиска по словарю ищет заданное слово и возвращает леммы и соответствующие им - части речи. Сокращения — соответствующие частям речи описаны в соответствующем языковом файле (/doc/tagsets/).

6. Модуль определения величин.

Функционал схож с модулем определения дат\чисел.

7. Вероятностное определение части речи.

На вход принимает предложение, на выходе вероятностные характеристики части речи для каждого слова (даже если слово отсутствует в словаре). Обучение проводилось по большому корпусу размеченных предложений со снятой омонимией. Используется классическая триграмная схема Маркова. На данный момент ошибка в определении краткой (часть речи — род) формы, не составляет больше 4.7%.