Автоматическая обработка текстов

СЕГМЕНТАЦИЯ

Понятия

Слово (токен) – единица графематического анализа

Словоформа – слово в определенной грамматической форме: *столе, столом*

Лексема – это слово как совокупность всех его конкретных грамматических форм (словоформ), обладающих тождественным лексическим значением: *стол*

Парадигма – список словоформ, принадлежащих одной лексеме {*стол, столе, столом, стола, столы, столами, …*}

Выделение предложения

Разделители предложений?

Выделение предложения

Разделители предложений?

Знаки пунктуации: . ! ?

Примеры:

- В этом году Московскому государственному техническому университету им. Н. Э. Баумана исполняется 175 лет.
- Умер Аменхотеп III примерно в 1367 г. до н. э.
- — Здоровы ли вы? — спросил Волков.
- — Какое здоровье! — зевая, сказал Обломов.
- https://petrsu.ru/news?dir=14

Выделение слова

Разделители слов?

Выделение слова

Разделители слов?

Пробел

Знаки пунктуации: , . ; ! : ? / – ( ) [ ] { } “”

Выделение слова

Разделители слов?

Пробел

Знаки пунктуации: , . ; ! : ? / – ( ) [ ] { } “”

Примеры:

- как будто, потому что, под рукой, в соответствии с, в общем и целом
- во-первых, по-летнему, вице-адмирал, как-то
- т. е., и т. д., до н. э.
- м/с, км/ч
- itsupport@petrsu.ru
- https://petrsu.ru/news?dir=14

Выделение слова

Разделители слов?

Пробел

Знаки пунктуации: , . ; ! : ? / – ( ) [ ] { } “”

Примеры:

- Lebensversicherungsgesellschaftsangestellter – сотрудник компании по страхованию жизни
- Donaudampfschifffahrtskapitän – капитан рейса парохода по Дунаю
- 經過132年的法國占領，阿爾及利亞於1962年獲得獨立
- نالت الجزائر استقلالها عام 1962 بعد 132 عامًا من الاحتلال الفرنسي

Выделение слова

Важны ли:

- Регистр букв

Примеры

- БАМ – Байкало-Амурская магистраль
- Бам — город на юго-востоке Ирана
- Бам — остров в Тихом океане
- бам – «Слышишь, время летит — бам!»
- Гончаров и гончаров

Выделение слова

Важны ли:

- Цифры

Примеры

- 80-летие, 5-местный, 6-й, МиГ-29, 10 млн., Уран-235
- 3904,78 (3904.78, 3,904.78), 1 1/2, 45°, –10,6 °C, +10 °C ,
- 10–20 тыс. м, 20 %, 20 %-го,
- 20.01.2021, 61°47′46″ с. ш.
- Даты: 01.01.2021 г. (01/01/2021)
- Номер телефона: +7 123 123–45–67, +7 (123) 123–45–67

Выделение слова

Важны ли:

- Вертикальное смещение

Примеры

- CH_3CH_2OH
- $f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$

Задача

1) Доли и доли: доли в общем деле, и доли в общей судьбе.
2) Объективный в общем-то процесс.
3) В общем, картина была такова.
4) Это и есть в самом общем виде задача правозащитников, не так ли?
5) В самом общем виде зависимость от интернета проявляется в том, что люди предпочитают "виртуальную" жизнь "реальной".
6) В общем списке была ловушка под названием "Бесплатное богатство".

Сколько слов?

Сколько словоформ?

Сколько лексем?