Методы обработки текстов на естественном языке

Задачи обработки текстов

- Частеречная разметка (part-of-speech tagging): разметить в заданном тексте слова по частям речи и, возможно, по морфологическим признакам;
- Морфологическая сегментация (morphological segmentation): разделить слова на морфемы (приставки, суффиксы и т.д.);
- Языковые модели (language models): по заданному отрывку текста предсказать следующее слово или символ;
- Анализ тональности (sentiment analysis): определить позитивное или негативное отношение несет текст;
- Выделение отношений или фактов (relationship extraction): выделить из текста факты об упоминающихся там сущностях;

Задачи обработки текстов

- Ответы на вопросы (question answering): дать ответ на заданный вопрос. Классификация или порождение текста (поддержание диалога). Задачи на здравый смысл: «Кубок не помещался в чемодан, потому что он был слишком велик; что именно было слишком велико, чемодан или кубок?»
- Порождение текста (text generation);
- Информационный поиск (information retrieval): найти релевантные документы;
- Машинный перевод (machine translation): по тексту на одном языке породить соответствующий текст на другом языке;
- Диалоговые модели (dialog and conversational models): поддержать разговор с человеком – чат-боты.

Мешок слов

Дана коллекция **текстовых документов** $D = \{d_1, \dots, d_n\}$.

Каждый документ характеризуется **набором слов** $W_d = \{w_1, \dots, w_{m_d}\}$ в некотором порядке.

Все возможные слова во всех документах образуют **корпус слов** V.

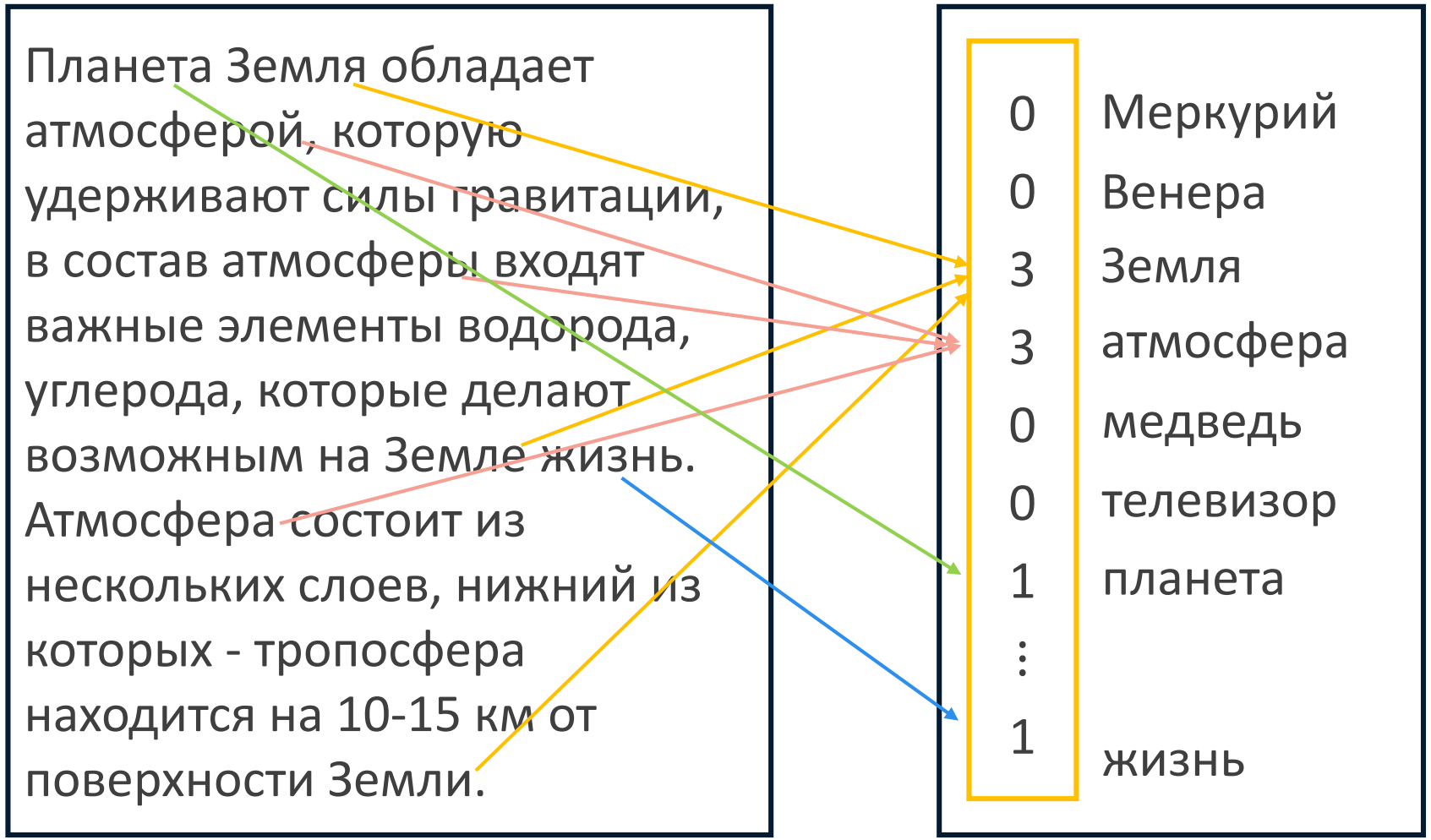
Если порядок слов неважен, то каждый документ представляет собой **мешок слов** (*bag of words*).

Каждый документ можно охарактеризовать по тем словам, которые в него входят, посчитав число вхождений каждого слова из корпуса:

$$d_i \longmapsto \left(n_{i,w_1}, n_{i,w_2}, \dots, n_{i,w_{|V|}}\right),$$

где $|V|$ – число слов в корпусе.

Мешок слов


Планета Земля обладает атмосферой, которую удерживают силы гравитации, в состав атмосферы входят важные элементы водорода, углерода, которые делают возможным на Земле жизнь. Атмосфера состоит из нескольких слоев, нижний из которых - тропосфера находится на 10-15 км от поверхности Земли.
0 Меркурий
0 Венера
3 Земля
3 атмосфера
0 медведь
0 телевизор
1 планета
⋮
1 жизнь

Нормализация текста

Стемминг – нахождение основы слова для заданного исходного слова.

лесной → лес

походный → поход

столовый → стол

Лемматизация – приведение слова к нормальной (словарной) форме.

бежала → бежать

кошку → кошка

зеленого → зеленый

TF-IDF

Идея метода TF-IDF: вычислять оценки важности слов для документа:

- чем чаще слово встречается в документе, тем оно важнее;
- чем реже слово встречается в других документах, тем оно важнее.

$n_{i,w}$ (term frequency) – число вхождений слова w в текст d_i;

N_w (document frequency) – число текстов, содержащих w.

$$\text{TF-IDF}(i,w) = n_{i,w} \times \log\left(\frac{N}{N_w}\right)$$

$\text{TF}(i,w) = n_{i,w}$ – term frequency;

$\text{IDF}(w) = \log(N/N_w)$ – inverted document frequency.

N-граммы

В «мешке слов» теряются связи между словами. Например, сочетания

«пошел на пару» и *«не пошел на пару»*,
«Новгород» и *«Нижний Новгород»* имеют разный смысл.

В таких случаях в корпус слов добавляются не только отдельные слова, но и все возможные пары слов (биграммы).

Биграммы:

«мама мыла раму» → (мама, мыла), (мыла раму)

Триграммы:

«Счастье есть удовольствие без раскаяния» → (счастье, есть, удовольствие), (есть, удовольствие, без), (удовольствие, без, раскаяния)