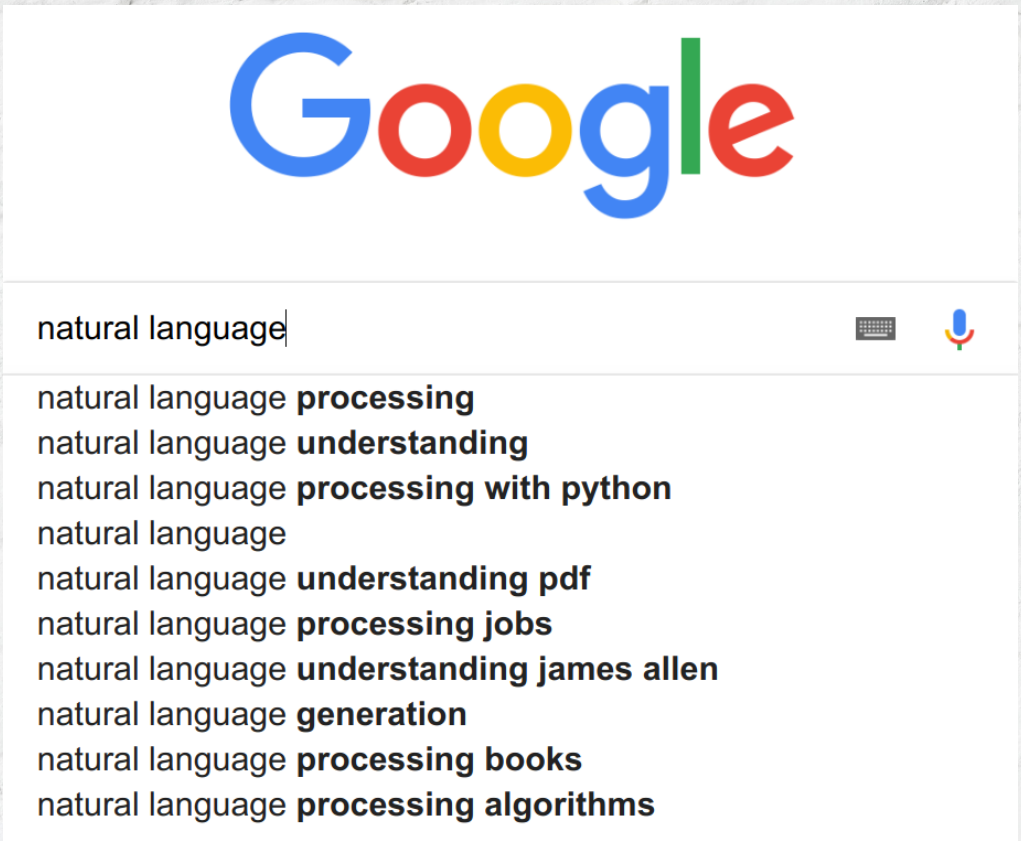
n-grams & markov chains



André Santos, afs@inesctec.pt

Counting frequencies

“Tudo aceitar, o que vem e o que foge, com a tranquilidade com que se acolhem as naturais mudanças de dias agrestes e de dias suaves. ”

Single words (unigrams)

```
que       : 3,    aceitar       : 1,    as          : 1,
o         : 2,    vem           : 1,    naturais    : 1,
e         : 2,    foge          : 1,    mudanças    : 1,
com       : 2,    a             : 1,    agrestes    : 1,
de        : 2,    tranquilidade : 1,    suaves      : 1,
dias      : 2,    se            : 1,    tudo        : 1,
acolhem   : 1
```

Bigrams

“Tudo aceitar, o que vem e o que foge, com a tranquilidade com que se acolhem as naturais mudanças de dias agrestes e de dias suaves.”

Occurences of two words

```
de dias           : 2,    as naturais       : 1,    foge com    : 1,
o que             : 2,    acolhem as        : 1,    que foge    : 1,
que se            : 1,    se acolhem        : 1,    e o         : 1,
e de              : 1,    com que           : 1,    vem e       : 1,
agrestes e        : 1,    tudo aceitar      : 1,    que vem     : 1,
dias agrestes     : 1,    tranquilidade com : 1,    aceitar o   : 1,
mudanças de       : 1,    a tranquilidade   : 1,    dias suaves : 1,
naturais mudanças : 1,    com a             : 1
```

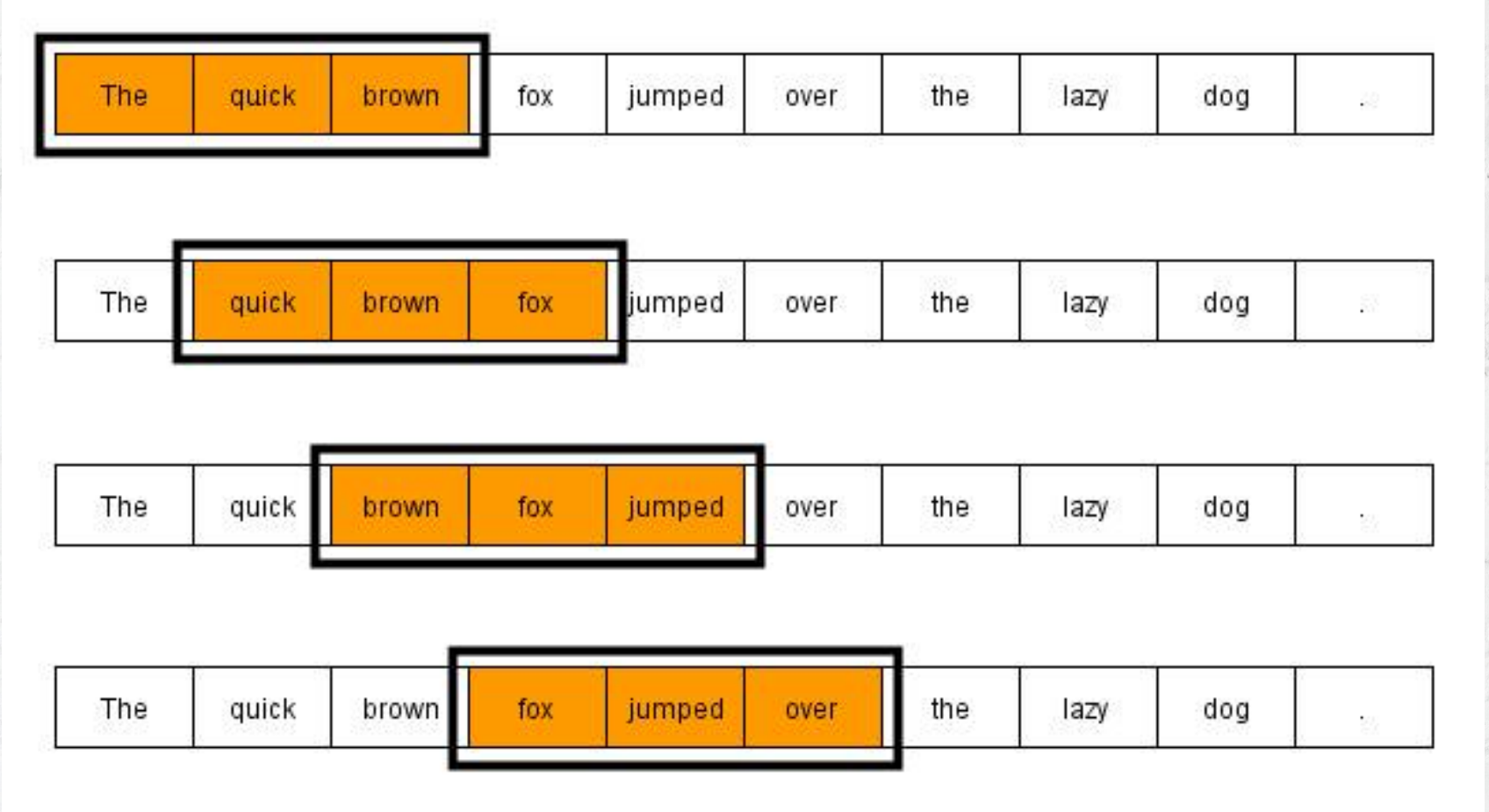
Trigrams

“Tudo aceitar, o que vem e o que foge, com a tranquilidade com que se acolhem as naturais mudanças de dias agrestes e de dias suaves.”

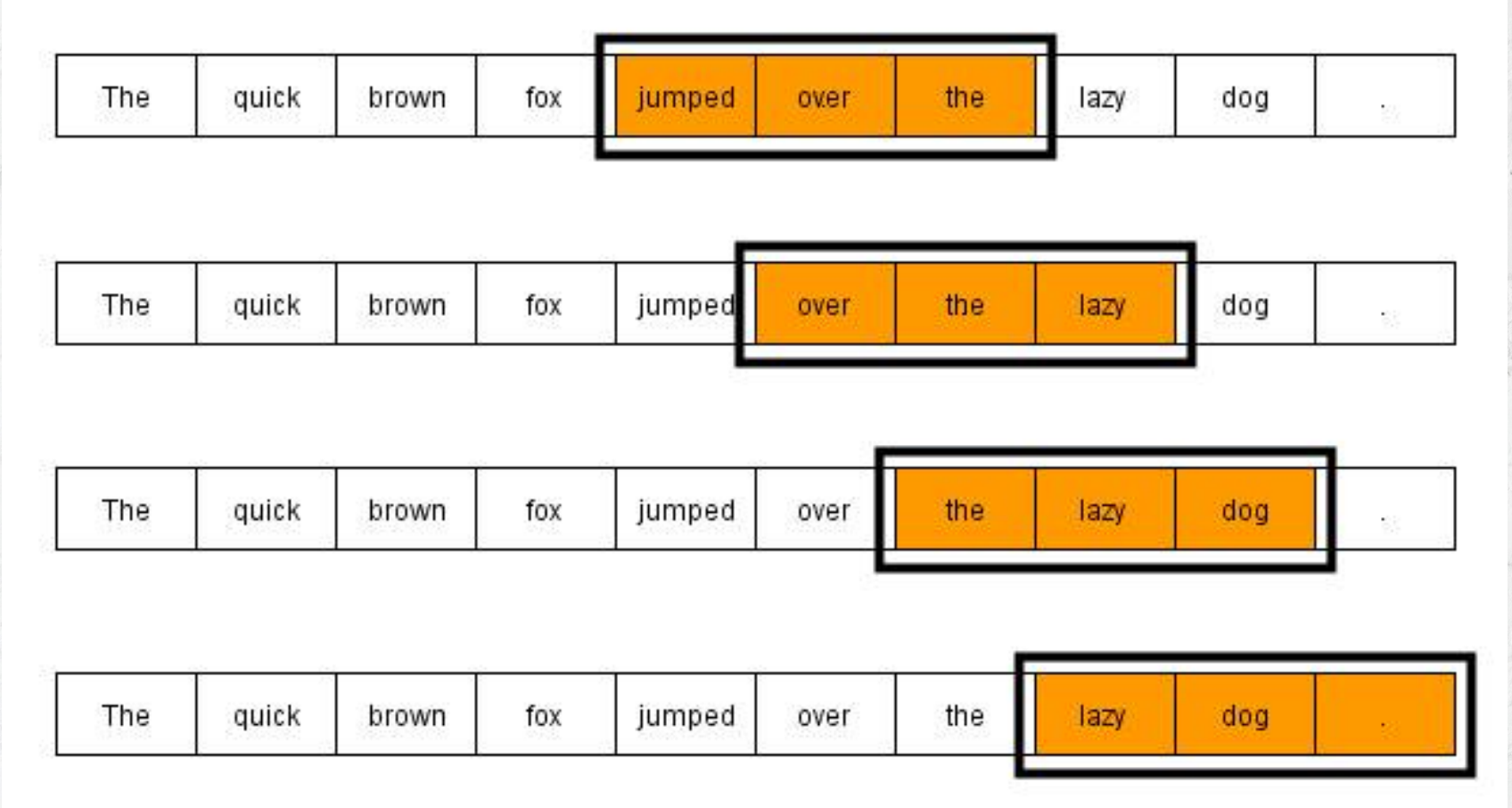
Occurences of three words

```
com que se            : 1, acolhem as naturais   : 1, que foge com          : 1,
e de dias             : 1, se acolhem as         : 1, o que foge           : 1,
agrestes e de         : 1, que se acolhem        : 1, e o que              : 1,
dias agrestes e       : 1, tranquilidade com que : 1, vem e o              : 1,
de dias agrestes      : 1, Tudo aceitar          : 1, que vem e            : 1,
mudanças de dias      : 1, a tranquilidade com   : 1, o que vem            : 1,
naturais mudanças de  : 1, com a tranquilidade   : 1, aceitar o que        : 1,
as naturais mudanças  : 1, foge com a            : 1, tudo aceitar o       : 1
```

Trigrams


The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .

Trigrams


The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .
The quick brown fox jumped over the lazy dog .

N-grams

Sequence of N words (or characters)

Applications

- determine the likelihood of an automated machine translation being correct
- predict the next most likely word to occur in a sentence,
- automatically generate text from speech,
- automate spelling correction,
- determine the relative sentiment of a piece of text.

Conditional probability

If X and Y are two events, then the conditional probability of X w.r.t. Y is denoted

$$P(X \mid Y)$$

The probability of event X given that Y has already occurred.

Chain rule of probability

$$\mathbf{P}(\mathbf{A_n}, \ldots, \mathbf{A}_1) = \mathbf{P}(\mathbf{A_n} | \mathbf{A_n}_{-1}, \ldots, \mathbf{A}_1) \cdot \mathbf{P}(\mathbf{A_n}_{-1}, \ldots, \mathbf{A}_1)$$

$$\begin{aligned}\mathbf{P}(\mathbf{A}_4, \mathbf{A}_3, \mathbf{A}_2, \mathbf{A}_1) = \; & \mathbf{P}(\mathbf{A}_4 \mid \mathbf{A}_3, \mathbf{A}_2, \mathbf{A}_1) \cdot \\ & \mathbf{P}(\mathbf{A}_3 \mid \mathbf{A}_2, \mathbf{A}_1) \cdot \\ & \mathbf{P}(\mathbf{A}_2 \mid \mathbf{A}_1) \cdot \\ & \mathbf{P}(\mathbf{A}_1)\end{aligned}$$

Chain rule of probability

$$\mathbf{P}(\mathbf{Ainda}, \mathbf{o}, \mathbf{apanhamos}) = \mathbf{P}(\mathbf{apanhamos} \mid \mathbf{Ainda}, \mathbf{o}) \cdot \mathbf{P}(\mathbf{o} \mid \mathbf{Ainda}) \cdot \mathbf{P}(\mathbf{Ainda})$$

$$\mathbf{P}\left(\bigcap_{\mathbf{k}=1}^{\mathbf{n}} \mathbf{A_k}\right) = \prod_{\mathbf{k}=1}^{\mathbf{n}} \mathbf{P}\left(\mathbf{A_k} \,\middle|\, \bigcap_{\mathbf{j}=1}^{\mathbf{k}-1} \mathbf{A_j}\right)$$

Markov assumption

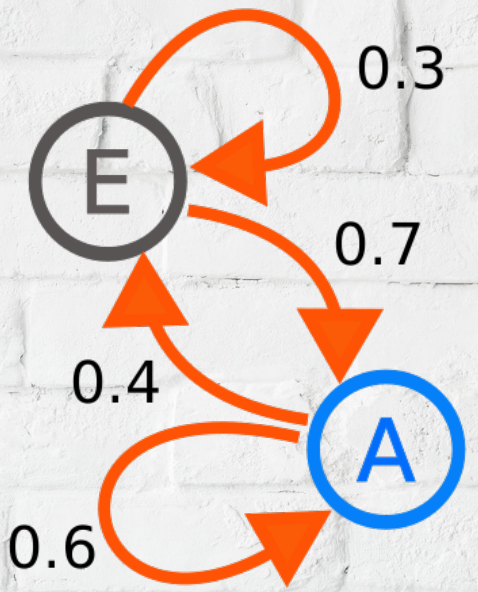
The probability of any given event can be approximated by taking into account only the closest past events.

$$\mathbf{P}(\mathbf{A_n}|\mathbf{A_{n-1}}, \ldots, \mathbf{A}_1) \sim \mathbf{P}(\mathbf{A_n}|\mathbf{A_{n-1}})$$

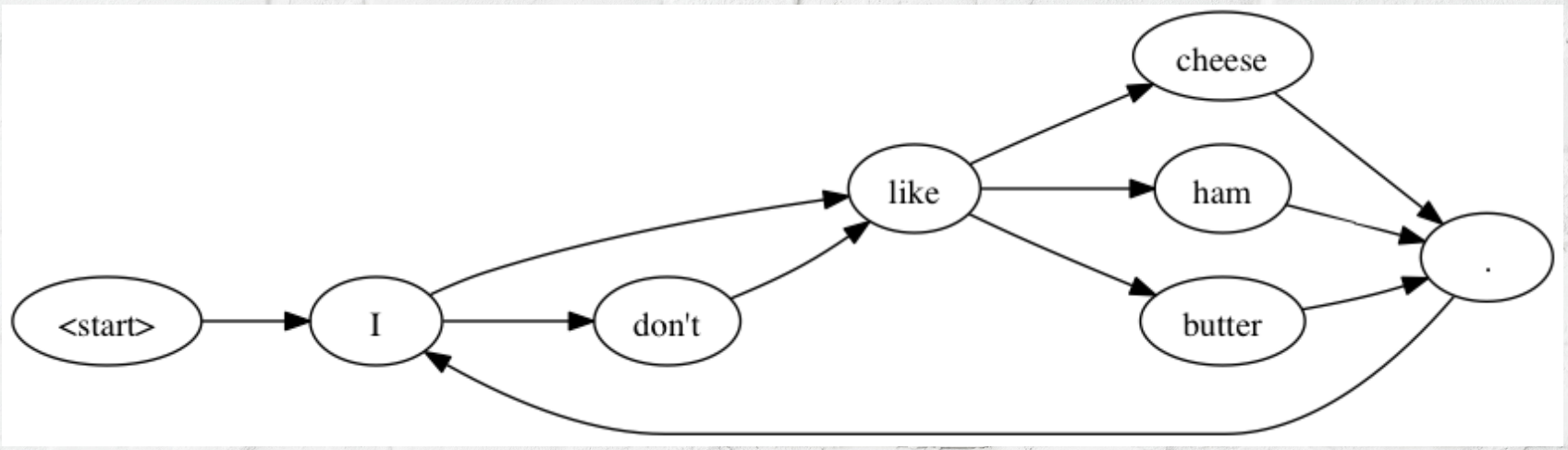
$$\mathbf{P}(\mathbf{apanhamos} \mid \mathbf{Ainda}, \mathbf{o}) \sim \mathbf{P}(\mathbf{apanhamos} \mid \mathbf{o})$$

Markov chain

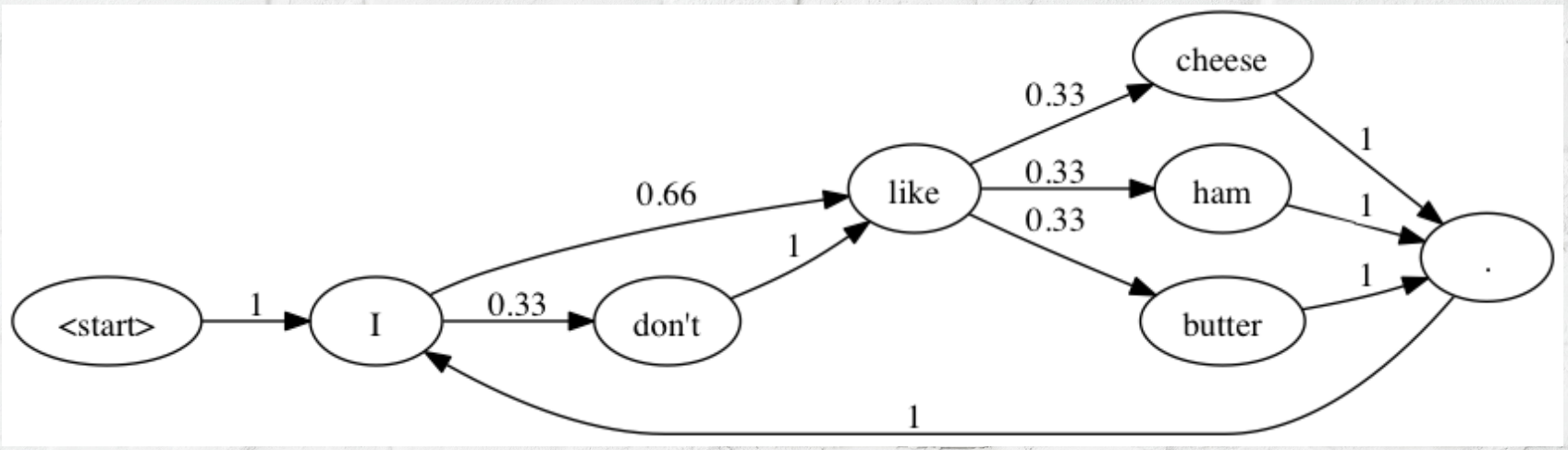
A sequence of random "states" where each new state is conditional only on the previous state.



Markov chain



Markov chain



Exercises

1. Implement a function which, given a text, generates sentences by applying a Markov chain to that text's N-grams. Watch what happens when you increase/decrease N.
2. Implement a function which, given a large text, uses character N-grams to correct typos in a second (smaller) text.

Python: Counter

dict subclass for counting hashable objects

`https://docs.python.org/3/library/collections.html#collections.Counter`