Contagens Simples em Corpora

Prof. Marcos Lopes

Departamento de Linguística – USP

Tycho Brahe (Svalöv, 1546 – Praga, 1601)

Corpora Não Estruturados

Lei de Zipf

Riqueza Lexical

Riqueza Vocabular

Hápax Legômena

Corpora Não Estruturados

- Um corpus não estruturado é aquele cujos dados não estão previamente *categorizados*, ou seja, não estão divididos em *campos* como as colunas de uma tabela, e tampouco estão *anotados* (com etiquetas morfossintáticas, por exemplo).

Corpora Não Estruturados

- Um corpus não estruturado é aquele cujos dados não estão previamente *categorizados*, ou seja, não estão divididos em *campos* como as colunas de uma tabela, e tampouco estão *anotados* (com etiquetas morfossintáticas, por exemplo).
- Tipicamente, são arquivos de texto bruto (extensão .TXT), sem formatação, ou conteúdos textuais extraídos de páginas da internet, mensagens curtas de texto e coisas assim.

Corpora Não Estruturados

- Um corpus não estruturado é aquele cujos dados não estão previamente *categorizados*, ou seja, não estão divididos em *campos* como as colunas de uma tabela, e tampouco estão *anotados* (com etiquetas morfossintáticas, por exemplo).
- Tipicamente, são arquivos de texto bruto (extensão .TXT), sem formatação, ou conteúdos textuais extraídos de páginas da internet, mensagens curtas de texto e coisas assim.
- Isso significa que o conteúdo é sujo (com caracteres estranhos no meio do texto, por exemplo) e heterogêneo (nem tudo é conteúdo ali: pode haver metadados das páginas web, informações editoriais sobre o texto etc.).

Corpora Não Estruturados

- Um corpus não estruturado é aquele cujos dados não estão previamente *categorizados*, ou seja, não estão divididos em *campos* como as colunas de uma tabela, e tampouco estão *anotados* (com etiquetas morfossintáticas, por exemplo).
- Tipicamente, são arquivos de texto bruto (extensão .TXT), sem formatação, ou conteúdos textuais extraídos de páginas da internet, mensagens curtas de texto e coisas assim.
- Isso significa que o conteúdo é sujo (com caracteres estranhos no meio do texto, por exemplo) e heterogêneo (nem tudo é conteúdo ali: pode haver metadados das páginas web, informações editoriais sobre o texto etc.).
- Foge aos propósitos deste nosso curso entrar em detalhes sobre a limpeza desses dados, mas lembre-se: numa tarefa real, você terá de limpá-los.

Leitura de arquivos TXT

No Python, pode-se abrir um arquivo de texto bruto para leitura ('r') ou escrita ('w').

```
arq = open('Guarani.txt', 'r')
texto = arq.read()
arq.close()
```

É possível percorrer o arquivo linha por linha. Observe que, no caso a seguir, a codificação de caracteres também foi passada como parâmetro à função **open**():

```
arquivo = open('Guarani.txt', 'r', encoding='utf-8')
for linha in arquivo:
        print(linha)
arquivo.close()
```

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)
  - Existe um cabeçalho?

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)
  - Existe um cabeçalho?
  - Existem metadados de algum tipo?

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)
  - Existe um cabeçalho?
  - Existem metadados de algum tipo?
- O corpus representa tudo o que se tem como material para análise. Não se pode acrescentar nada.

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)
  - Existe um cabeçalho?
  - Existem metadados de algum tipo?
- O corpus representa tudo o que se tem como material para análise. Não se pode acrescentar nada.
- É possível filtrar dados **desde que não manualmente**. Tem de haver um algoritmo ou, ao menos, um critério transparente de filtragem.

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)
  - Existe um cabeçalho?
  - Existem metadados de algum tipo?
- O corpus representa tudo o que se tem como material para análise. Não se pode acrescentar nada.
- É possível filtrar dados **desde que não manualmente**. Tem de haver um algoritmo ou, ao menos, um critério transparente de filtragem.
- Se aplicados, os filtros têm de ser apresentados explicitamente na publicação do código, na seção de Métodos de um trabalho acadêmico etc.

A observar atentamente

- Antes de começar a processar os dados, procure abrir o arquivo (ou um dos arquivos) do corpus e conhecer os dados.
- Observe:
  - A codificação do arquivo (UTF-8? ISO?...)
  - Existe um cabeçalho?
  - Existem metadados de algum tipo?
- O corpus representa tudo o que se tem como material para análise. Não se pode acrescentar nada.
- É possível filtrar dados **desde que não manualmente**. Tem de haver um algoritmo ou, ao menos, um critério transparente de filtragem.
- Se aplicados, os filtros têm de ser apresentados explicitamente na publicação do código, na seção de Métodos de um trabalho acadêmico etc.
- **Jamais manipule diretamente os arquivos originais**. Faça cópias para suas análises.

Corpora Não Estruturados

Lei de Zipf

Riqueza Lexical

Riqueza Vocabular

Hápax Legômena

Lei de Zipf

- O linguista estadunidense George Zipf (1902 – 1950) observou a distribuição da frequência de palavras em corpora textuais do tamanho de livros.

Lei de Zipf

- O linguista estadunidense George Zipf (1902 – 1950) observou a distribuição da frequência de palavras em corpora textuais do tamanho de livros.
    - O tamanho de um livro pode parecer modesto, mas já era um desafio considerável quando não havia computadores.

Lei de Zipf

- O linguista estadunidense George Zipf (1902 – 1950) observou a distribuição da frequência de palavras em corpora textuais do tamanho de livros.
    - O tamanho de um livro pode parecer modesto, mas já era um desafio considerável quando não havia computadores.
- Analisando textos de língua inglesa, ele estabeleceu:

Lei de Zipf

- O linguista estadunidense George Zipf (1902 – 1950) observou a distribuição da frequência de palavras em corpora textuais do tamanho de livros.
    - O tamanho de um livro pode parecer modesto, mas já era um desafio considerável quando não havia computadores.
- Analisando textos de língua inglesa, ele estabeleceu:
    - O enunciado geral da lei que leva seu nome;

Lei de Zipf

- O linguista estadunidense George Zipf (1902 – 1950) observou a distribuição da frequência de palavras em corpora textuais do tamanho de livros.
    - O tamanho de um livro pode parecer modesto, mas já era um desafio considerável quando não havia computadores.
- Analisando textos de língua inglesa, ele estabeleceu:
    - O enunciado geral da lei que leva seu nome;
    - Uma constante da distribuição relativa aos textos em inglês.

Lei de Zipf (cont.)

Lei geral de Zipf

A frequência f de um evento é inversamente proporcional à sua ordem em uma classificação (*ranking*, r).

Lei de Zipf (cont.)

Lei geral de Zipf

A frequência f de um evento é inversamente proporcional à sua ordem em uma classificação (*ranking*, r).

Lei de Zipf (cont.)

Lei geral de Zipf

A frequência f de um evento é inversamente proporcional à sua ordem em uma classificação (*ranking*, r).

Para os textos de língua inglesa, Zipf mostrou que:

$$f(r) \simeq \frac{0,1}{r}$$

Lei de Zipf (cont.)

Lei geral de Zipf

A frequência f de um evento é inversamente proporcional à sua ordem em uma classificação (*ranking*, r).

Para os textos de língua inglesa, Zipf mostrou que:

$$f(r) \simeq \frac{0,1}{r}$$

- Ou seja, a palavra mais frequente ($r = 1$) ocorre $\frac{0,1}{1} = 0,1$, isto é, 10% das vezes.

Lei de Zipf (cont.)

Lei geral de Zipf

A frequência f de um evento é inversamente proporcional à sua ordem em uma classificação (*ranking*, r).

Para os textos de língua inglesa, Zipf mostrou que:

$$f(r) \simeq \frac{0,1}{r}$$

- Ou seja, a palavra mais frequente ($r = 1$) ocorre $\frac{0,1}{1} = 0,1$, isto é, 10% das vezes.
- A segunda palavra mais frequente ($r = 2$) ocorre $\frac{0,1}{2} = 0,05$, isto é, 5% das vezes, e assim por diante.

Lei de Zipf (cont.)

Lei geral de Zipf

A frequência f de um evento é inversamente proporcional à sua ordem em uma classificação (*ranking*, r).

Para os textos de língua inglesa, Zipf mostrou que:

$$f(r) \cong \frac{0,1}{r}$$

- Ou seja, a palavra mais frequente ($r = 1$) ocorre $\frac{0,1}{1} = 0,1$, isto é, 10% das vezes.
- A segunda palavra mais frequente ($r = 2$) ocorre $\frac{0,1}{2} = 0,05$, isto é, 5% das vezes, e assim por diante.
- A lei não se aplica além de $r = 1.000$.

Corpora Não Estruturados

Lei de Zipf

Riqueza Lexical

Riqueza Vocabular

Hápax Legômena

Riqueza lexical

- A riqueza lexical de um documento é dada pela razão da diversidade do repertório de palavras usado pela extensão do documento.

Riqueza lexical

- A riqueza lexical de um documento é dada pela razão da diversidade do repertório de palavras usado pela extensão do documento.
- Quanto mais formas distintas, mais “rico” o documento.

Riqueza lexical

- A riqueza lexical de um documento é dada pela razão da diversidade do repertório de palavras usado pela extensão do documento.
- Quanto mais formas distintas, mais “rico” o documento.
- No Python, o número de formas distintas é calculado simplesmente transformando a lista de tokens em um conjunto de tokens, o que elimina os duplicados.

Riqueza lexical

- A riqueza lexical de um documento é dada pela razão da diversidade do repertório de palavras usado pela extensão do documento.
- Quanto mais formas distintas, mais “rico” o documento.
- No Python, o número de formas distintas é calculado simplesmente transformando a lista de tokens em um conjunto de tokens, o que elimina os duplicados.
- Em seguida, esse número é dividido pelo total de tokens.

Corpora Não Estruturados

Lei de Zipf

Riqueza Lexical

Riqueza Vocabular

Hápax Legômena

Riqueza vocabular

- O vocabulário de um documento é o conjunto de seus lemas.

Riqueza vocabular

- O vocabulário de um documento é o conjunto de seus lemas.
- Essa definição é, muitas vezes, simplificada para “conjunto dos tokens sem duplicação”.

Riqueza vocabular

- O vocabulário de um documento é o conjunto de seus lemas.
- Essa definição é, muitas vezes, simplificada para “conjunto dos tokens sem duplicação”.
- Alguns autores distinguem a riqueza lexical da riqueza vocabular, que seria calculada não com base nas palavras do documento, mas no vocabulário (nos lemas).

Riqueza vocabular

- O vocabulário de um documento é o conjunto de seus lemas.
- Essa definição é, muitas vezes, simplificada para “conjunto dos tokens sem duplicação”.
- Alguns autores distinguem a riqueza lexical da riqueza vocabular, que seria calculada não com base nas palavras do documento, mas no vocabulário (nos lemas).
- Só é possível fazer isso, claro, se você dispuser de um lematizador razoável. No caso do português, isso não existia até 2015, quando surgiu o spaCy.

Riqueza vocabular

- O vocabulário de um documento é o conjunto de seus lemas.
- Essa definição é, muitas vezes, simplificada para “conjunto dos tokens sem duplicação”.
- Alguns autores distinguem a riqueza lexical da riqueza vocabular, que seria calculada não com base nas palavras do documento, mas no vocabulário (nos lemas).
- Só é possível fazer isso, claro, se você dispuser de um lematizador razoável. No caso do português, isso não existia até 2015, quando surgiu o spaCy.
- Por conta dessa restrição técnica, a riqueza lexical é mais presente na literatura especializada que a vocabular.

Corpora Não Estruturados

Lei de Zipf

Riqueza Lexical

Riqueza Vocabular

Hápax Legômena

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:
    - Raridade da palavra

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:
  - Raridade da palavra
  - Raridade da forma realizada (da flexão, da junção com pronomes hifenizados)

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:
  - Raridade da palavra
  - Raridade da forma realizada (da flexão, da junção com pronomes hifenizados)
  - Estrangeirismos

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:
  - Raridade da palavra
  - Raridade da forma realizada (da flexão, da junção com pronomes hifenizados)
  - Estrangeirismos
  - Neologismos

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:
  - Raridade da palavra
  - Raridade da forma realizada (da flexão, da junção com pronomes hifenizados)
  - Estrangeirismos
  - Neologismos
  - Erros de digitação ou ortográficos

Απαξ λεγομενα

- *Hápax legômenon* (plural: *hápax legômena*) é o termo técnico usado para designar as palavras com uma única ocorrência no corpus sob análise.
- Conhecer os hápax legômena tem utilidade prática muito distinta daquela das palavras frequentes.
- Há diferentes razões para uma palavra ocorrer uma única vez:
  - Raridade da palavra
  - Raridade da forma realizada (da flexão, da junção com pronomes hifenizados)
  - Estrangeirismos
  - Neologismos
  - Erros de digitação ou ortográficos
- São muitas as análises que decidem deixar de fora os hápax legômena. Além das razões acima, a principal motivação para isso é reduzir a dimensionalidade nas abordagens baseadas em