Modelos de n-gramas

Marcos Lopes

Departamento de Linguística

Modelos de n-gramas

Implementação

Encadeamentos lineares

- Modelos BoW têm limites práticos

Encadeamentos lineares

- Modelos BoW têm limites práticos
- Existem problemas nas expressões linguísticas que só aparecem nos encadeamentos

Encadeamentos lineares

- Modelos BoW têm limites práticos
- Existem problemas nas expressões linguísticas que só aparecem nos encadeamentos
- Os exemplos a seguir não apresentam erros em palavras, mas encadeamentos inesperados:

Encadeamentos lineares

- Modelos BoW têm limites práticos
- Existem problemas nas expressões linguísticas que só aparecem nos encadeamentos
- Os exemplos a seguir não apresentam erros em palavras, mas encadeamentos inesperados:
    - Praia Clube é treta campeão mineiro de vôlei.

Encadeamentos lineares

- Modelos BoW têm limites práticos
- Existem problemas nas expressões linguísticas que só aparecem nos encadeamentos
- Os exemplos a seguir não apresentam erros em palavras, mas encadeamentos inesperados:
    - Praia Clube é treta campeão mineiro de vôlei.
    - O tráfego de drogas faz muitas vítimas entre os jovens.

Encadeamentos lineares

- Modelos BoW têm limites práticos
- Existem problemas nas expressões linguísticas que só aparecem nos encadeamentos
- Os exemplos a seguir não apresentam erros em palavras, mas encadeamentos inesperados:
  - Praia Clube é treta campeão mineiro de vôlei.
  - O tráfego de drogas faz muitas vítimas entre os jovens.
  - Esta mesa tem dois metros de cumprimento.

Alguns usos

- Corretores ortográticos e sintáticos

Alguns usos

- Corretores ortográticos e sintáticos
- Tradutores automáticos

Alguns usos

- Corretores ortográticos e sintáticos
- Tradutores automáticos
- Reconhecimento de voz

Alguns usos

- Corretores ortográticos e sintáticos
- Tradutores automáticos
- Reconhecimento de voz
- POS-tagging

Encadeamentos de unidades independentes

- O modelo mais simples possível é aquele que associa uma frequência relativa à unidade (à palavra, por ex.) e faz equivaler a chance de ocorrência a essa frequência.

Encadeamentos de unidades independentes

- O modelo mais simples possível é aquele que associa uma frequência relativa à unidade (à palavra, por ex.) e faz equivaler a chance de ocorrência a essa frequência.
- Por ex., se a língua conta com 500.000 palavras-tipo, a probabilidade de ocorrência de uma palavra qualquer em uma sequência qualquer seria de $\frac{1}{500.000}$ (0,000002 ou 0,0002%).

Encadeamentos de unidades independentes

- O modelo mais simples possível é aquele que associa uma frequência relativa à unidade (à palavra, por ex.) e faz equivaler a chance de ocorrência a essa frequência.
- Por ex., se a língua conta com 500.000 palavras-tipo, a probabilidade de ocorrência de uma palavra qualquer em uma sequência qualquer seria de $\frac{1}{500.000}$ (0,000002 ou 0,0002%).
- Essa solução pode ser adequada para eventos completamente independentes entre si (por ex., dois lançamentos de dados), mas esse não é o caso de *nenhuma* unidade linguística, sejam sílabas, palavras ou até respostas a perguntas (frases inteiras).

Tendências

Numa oração como:

O árbitro favoreceu o *x*

x sofre coerções lexicais, gramaticais e até discursivas. Algumas expressões são tendencialmente muito mais prováveis que outras.

Assim, a probabilidade de se ter uma palavra *x* como:

- *algumas*: menor que $\frac{1}{500.000}$

Tendências

Numa oração como:

O árbitro favoreceu o *x*

x sofre coerções lexicais, gramaticais e até discursivas. Algumas expressões são tendencialmente muito mais prováveis que outras.

Assim, a probabilidade de se ter uma palavra *x* como:

- *algumas*: menor que $\frac{1}{500.000}$
- *Flamengo*: praticamente 100%

Probabilidade condicional

Seria possível calcular num corpus a probabilidade de *Flamengo* dado que *O árbitro favoreceu o*.

Pode ser difícil ou impossível (isto é, sem exemplos no corpus) calcular a probabilidade de ocorrência de um elemento em função de toda uma longa cadeia pregressa. É mais prático, por ora, pensar na probabilidade do encadeamento de só duas palavras: *o* e *Flamengo*. Vamos representá-la assim:

$$P(Flamengo \mid o)$$

Ou, generalizando:

$$P(w_n \mid w_{n-1})$$

N-gramas ou Cadeias de Markov

- São modelos em que a probabilidade de ocorrência do elemento w_n é dada em função dos elementos imediatamente precedentes w_{n-1}.

N-gramas ou Cadeias de Markov

- São modelos em que a probabilidade de ocorrência do elemento w_n é dada em função dos elementos imediatamente precedentes w_{n-1}.
- São modelos de memória limitada.

Modelo	n	Probabilidades
Quadrigramas	4	$P(vento \mid caminhando\ contra\ o)$
Trigramas	3	$P(vento \mid contra\ o)$
Bigramas	2	$P(vento \mid o)$
Unigramas	1	$P(vento)$

N-gramas ou Cadeias de Markov

- São modelos em que a probabilidade de ocorrência do elemento w_n é dada em função dos elementos imediatamente precedentes w_{n-1}.
- São modelos de memória limitada.

Modelo	n	Probabilidades
Quadrigramas	4	$P(vento \mid caminhando\ contra\ o)$
Trigramas	3	$P(vento \mid contra\ o)$
Bigramas	2	$P(vento \mid o)$
Unigramas	1	$P(vento)$

- A ideia é que é possível *aproximar* a probabilidade de toda a história pregressa da palavra (memória) usando somente as n palavras anteriores.

N-gramas ou Cadeias de Markov

- São modelos em que a probabilidade de ocorrência do elemento w_n é dada em função dos elementos imediatamente precedentes w_{n-1}.
- São modelos de memória limitada.

Modelo	n	Probabilidades
Quadrigramas	4	$P(vento \mid caminhando\ contra\ o)$
Trigramas	3	$P(vento \mid contra\ o)$
Bigramas	2	$P(vento \mid o)$
Unigramas	1	$P(vento)$

- A ideia é que é possível *aproximar* a probabilidade de toda a história pregressa da palavra (memória) usando somente as n palavras anteriores.
- A hipótese segundo a qual a probabilidade de uma palavra só depende da(s) anterior(es) é chamada *Hipótese de Markov*.

Alguns recursos para visualização de n-gramas

- Google N-gram Viewer

  `https://books.google.com/ngrams`

Alguns recursos para visualização de n-gramas

- Google N-gram Viewer

  `https://books.google.com/ngrams`

- Concordanciadores

  Como exemplo: AntConc

Modelos de n-gramas

Implementação

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.
    - É útil acrescentar marcadores de fronteira de sentença (início e fim de sentença). Os símbolos usuais são <s> e </s>.

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.
    - É útil acrescentar marcadores de fronteira de sentença (início e fim de sentença). Os símbolos usuais são <s> e </s>.
2. Escolha do modelo (uni-, bi-, trigramas...)

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.
    - É útil acrescentar marcadores de fronteira de sentença (início e fim de sentença). Os símbolos usuais são <s> e </s>.
2. Escolha do modelo (uni-, bi-, trigramas...)
3. Geração das cadeias de n-gramas.

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.
    - É útil acrescentar marcadores de fronteira de sentença (início e fim de sentença). Os símbolos usuais são <s> e </s>.
2. Escolha do modelo (uni-, bi-, trigramas...)
3. Geração das cadeias de n-gramas.
    - Será preciso gerar cadeias para todos os modelos de cadeia (superiores ao unigrama).

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.
    - É útil acrescentar marcadores de fronteira de sentença (início e fim de sentença). Os símbolos usuais são <s> e </s>.
2. Escolha do modelo (uni-, bi-, trigramas...)
3. Geração das cadeias de n-gramas.
    - Será preciso gerar cadeias para todos os modelos de cadeia (superiores ao unigrama).
    - Por ex., se você quer trabalhar com trigramas, deve gerar trigramas e bigramas.

Passos para a criação de modelos de n-gramas

1. Segmentação do corpus.
    - Se os n-gramas forem palavras (e não caracteres ou frases), o texto deve ser dividido primeiro em sentenças, depois em palavras.
    - No pré-processamento, você deve decidir se a pontuação deve ou não ser mantida.
    - É útil acrescentar marcadores de fronteira de sentença (início e fim de sentença). Os símbolos usuais são <s> e </s>.
2. Escolha do modelo (uni-, bi-, trigramas...)
3. Geração das cadeias de n-gramas.
    - Será preciso gerar cadeias para todos os modelos de cadeia (superiores ao unigrama).
    - Por ex., se você quer trabalhar com trigramas, deve gerar trigramas e bigramas.
4. Cálculo das probabilidades dos modelos.

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$
- Estimativa da probabilidade por MLE: normalização dos valores entre 0 e 1, dividindo-se a contagem de cada palavra pelo vocabulário V, isto é, dividindo-se o número de *types* pelo número de *tokens*.

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$
- Estimativa da probabilidade por MLE: normalização dos valores entre 0 e 1, dividindo-se a contagem de cada palavra pelo vocabulário V, isto é, dividindo-se o número de *types* pelo número de *tokens*.

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$
- Estimativa da probabilidade por MLE: normalização dos valores entre 0 e 1, dividindo-se a contagem de cada palavra pelo vocabulário V, isto é, dividindo-se o número de *types* pelo número de *tokens*.

Bigramas

- Marcação de início e fim de sentença *antes* da tokenização em palavras.

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$
- Estimativa da probabilidade por MLE: normalização dos valores entre 0 e 1, dividindo-se a contagem de cada palavra pelo vocabulário *V*, isto é, dividindo-se o número de *types* pelo número de *tokens*.

Bigramas

- Marcação de início e fim de sentença *antes* da tokenização em palavras.
- A probabilidade é dada pela divisão do número de ocorrências dos bigramas pelo número de unigramas da primeira palavra da cadeia:

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$
- Estimativa da probabilidade por MLE: normalização dos valores entre 0 e 1, dividindo-se a contagem de cada palavra pelo vocabulário *V*, isto é, dividindo-se o número de *types* pelo número de *tokens*.

Bigramas

- Marcação de início e fim de sentença *antes* da tokenização em palavras.
- A probabilidade é dada pela divisão do número de ocorrências dos bigramas pelo número de unigramas da primeira palavra da cadeia:

Cálculo das probabilidades dos n-gramas

Unigramas

- Contagem dos tokens: $C(w)$
- Estimativa da probabilidade por MLE: normalização dos valores entre 0 e 1, dividindo-se a contagem de cada palavra pelo vocabulário V, isto é, dividindo-se o número de *types* pelo número de *tokens*.

Bigramas

- Marcação de início e fim de sentença *antes* da tokenização em palavras.
- A probabilidade é dada pela divisão do número de ocorrências dos bigramas pelo número de unigramas da primeira palavra da cadeia:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$