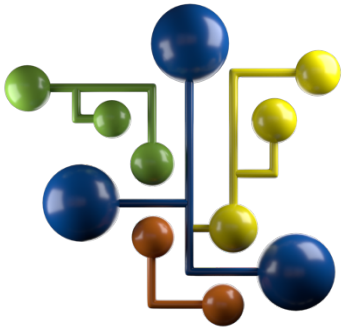

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Modelos de Palavras

Agora voltamos aos modelos de n-grama de palavras em vez de caracteres. Todos os mecanismos aplicam-se igualmente aos modelos de palavra e de caracteres. A principal diferença é que o vocabulário — o conjunto de símbolos que compõem o corpus e o modelo — é maior. Há apenas cerca de 100 caracteres na maioria dos idiomas e, às vezes, construímos modelos de caracteres que são ainda mais restritivos, por exemplo, tratando “A” e “a” como o mesmo símbolo ou tratando toda a pontuação como o mesmo símbolo. Mas, com modelos de palavras, temos pelo menos dezenas de milhares de símbolos e às vezes milhões.

A grande variedade é porque não é claro o que constitui uma palavra. Em inglês, uma sequência de letras rodeada por espaços é uma palavra, mas em algumas línguas, como o chinês, as palavras não são separadas por espaços, e até mesmo em inglês muitas decisões devem ser feitas para ter uma política clara sobre os limites da palavra: quantas palavras existem nestes 2 anagramas?

“ne'er-do-well”?

Ou em

“(Tel:1-800-960-5660x123)”?

Se é difícil para nós seres humanos, imagine para os computadores. Os modelos de palavras de n-grama têm de lidar com palavras fora do vocabulário. Com modelos de caracteres, não precisamos nos preocupar com alguém inventando uma nova letra no alfabeto. Mas, com modelos de palavras, há sempre a chance de uma nova palavra que não foi vista no corpus de treinamento, por isso precisamos modelar isso explicitamente em nosso modelo de linguagem.

Isso pode ser feito acrescentando apenas uma nova palavra ao vocabulário: <UNK>, como palavra desconhecida. Podemos estimar o resultado de n-grama para <UNK> através deste truque: passar pelo corpus de treinamento e, na primeira vez que qualquer palavra individual aparecer, sendo previamente desconhecida, substituímos pelo símbolo <UNK>. Todos os aparecimentos subsequentes da palavra permanecem inalterados. Então, calcula-se o resultado de n-grama para o corpus, como de costume, tratando <UNK> como qualquer outra palavra. Assim, quando uma palavra desconhecida aparecer em um conjunto de teste,

examinaremos a sua probabilidade sob <UNK>. Às vezes são utilizados símbolos de múltiplas palavras desconhecidas, para classes diferentes. Por exemplo, qualquer sequência de dígitos pode ser substituída por <NUM> ou qualquer endereço de e-mail, por <E-MAIL>.