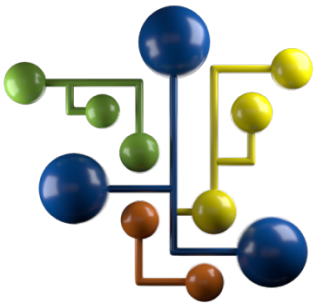

Data Science Academy

www.datascienceacademy.com.br

Business Analytics

O Que é Modelagem de Tópicos?

A maneira padrão de procurar documentos na internet é através de palavras-chave ou frases-chave. Isso é o que o Google e outros motores de busca fazem rotineiramente, e eles fazem isso muito bem. No entanto, por mais útil que seja, tem suas limitações. Considere, por exemplo, uma situação na qual você é confrontado com uma grande coleção de documentos, mas não tem ideia sobre o assunto que eles abordam. Uma das primeiras coisas que você pode querer fazer é classificar esses documentos em tópicos ou temas. Entre outras coisas isso iria ajudá-lo a descobrir o que há de interessante, enquanto também poderia orientá-lo sobre o subconjunto relevante dos dados. Para coleções pequenas, pode-se fazer isso simplesmente passando por cada documento, mas isso é claramente inviável para corpus contendo milhares de documentos.

A modelagem de tópicos – Topic Modeling - trata do problema de classificar automaticamente conjuntos de documentos em temas.

Topic Modeling é uma forma de mineração de texto, uma forma de identificar padrões em um corpus. Você agrupa palavras de todo o corpus em 'tópicos'. A modelagem de tópicos é "um método para encontrar e traçar clusters de palavras (chamados de "tópicos") em grandes corpus de textos".

O que, então, é um tópico? Uma definição oferecida no Twitter durante uma conferência sobre modelagem de tópicos descreveu um tópico como "um padrão recorrente de palavras co-ocorrentes". Uma ferramenta de modelagem de tópicos examina um corpus para esses grupos de palavras e os agrupa em conjunto por um processo de similaridade.

(ATENÇÃO: LDA também é a abreviatura de Linear Discriminant Analysis, uma técnica de classificação. Aqui estaremos falando de outro LDA, o Latent Dirichlet Allocation).

Latent Dirichlet Allocation

Em essência, LDA é uma técnica que facilita a descoberta automática de temas em uma coleção de documentos (corpus).

A suposição básica por trás de LDA é que cada um dos documentos em uma coleção consiste em uma mistura de tópicos de toda a coleção. No entanto, na realidade, observamos apenas documentos e palavras, não tópicos - estes últimos fazem parte da estrutura oculta (ou latente) dos documentos. O objetivo é inferir a estrutura tópica latente dada as palavras e o

documento. A LDA faz isso recriando os documentos no corpus e ajustando a importância relativa de tópicos em documentos e palavras em tópicos, iterativamente. O processo iterativo é implementado usando uma técnica chamada amostragem de Gibbs.

O termo "Dirichlet" em LDA refere-se ao fato de que tópicos e palavras são assumidos para seguir as distribuições de Dirichlet. Não há nenhuma "boa" razão para isso, além da conveniência - distribuições Dirichlet fornecem boas aproximações para distribuições de palavras em documentos e, talvez mais importante, são computacionalmente convenientes.

No link abaixo há uma descrição completa deste algoritmo, incluindo seu fundamento matemático:

http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Modelagem de Tópicos usando LDA

Agora estamos prontos para fazer alguma modelagem de tópicos. Usaremos o pacote topicmodels em R. Especificamente, usaremos a função LDA com a opção de amostragem Gibbs mencionada anteriormente. A função LDA tem um número bastante grande de parâmetros e usaremos valores padrões, embora você possa fazer um tunning dos parâmetros e customizar seus resultados.

A amostragem de Gibbs funciona executando uma caminhada aleatória de tal forma que reflete as características de uma distribuição desejada. Como o ponto de partida da caminhada é escolhido ao acaso, é necessário descartar os primeiros passos da caminhada (pois estes não refletem corretamente as propriedades da distribuição). Isto é referido como o período de burn-in.

Deve-se enfatizar que as configurações não garantem a convergência do algoritmo para uma solução globalmente ótima. Na verdade, a amostragem de Gibbs, na melhor das hipóteses, encontrará apenas uma solução localmente ótima, e mesmo isso é difícil de provar matematicamente em problemas práticos específicos, como o que estamos lidando aqui. O resultado disso é que é melhor fazer lotes de execuções com diferentes configurações de parâmetros para verificar a estabilidade de seus resultados. A questão é que nosso interesse é puramente prático por isso é bom o suficiente se os resultados fazem sentido.

Há um parâmetro importante que deve ser especificado inicialmente: k, o número de tópicos que o algoritmo deve usar para classificar documentos. Existem abordagens matemáticas para isso, mas muitas vezes não produzem escolhas semanticamente significativas de k. Do ponto de vista prático, pode-se simplesmente executar o algoritmo para diferentes valores de k e fazer uma escolha baseada na inspeção dos resultados. Isso é o que faremos.