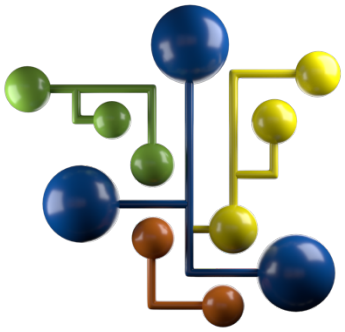


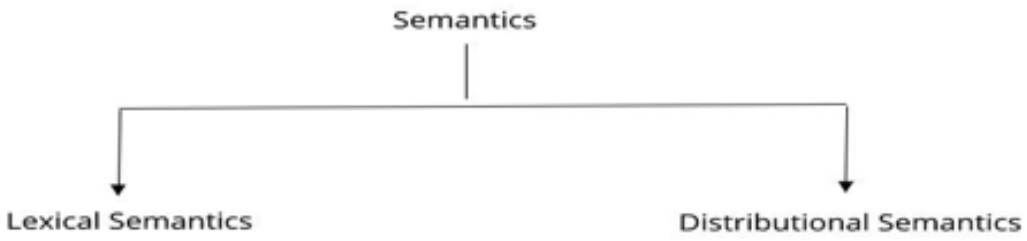
Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Semantica Distribucional

Nosso objetivo no capítulo anterior e agora neste capítulo, é lidar com a semântica no nível da palavra usando o Word2vec e em seguida, expandiremos nossos conceitos para nível de parágrafo e nível de documento. Observando a figura abaixo, você percebe os diferentes tipos de semânticas com os quais teremos que lidar:



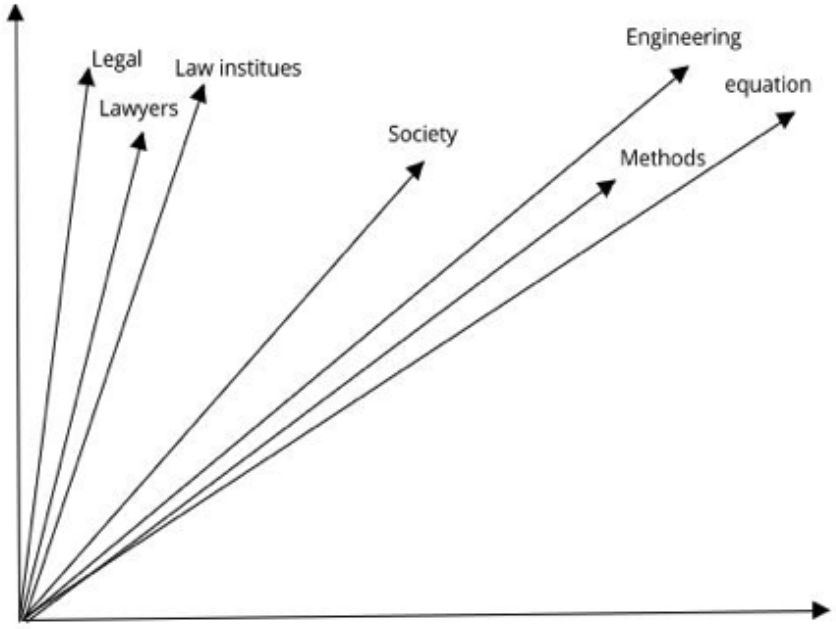
A semântica é um ramo que lida com o significado em PLN. Já cobrimos a semântica léxica nos capítulos anteriores, entendendo a estrutura das sentenças e agora, estamos estudando a semântica distribucional, área da qual Word2vec faz parte. Há também outras técnicas ou tipos na semântica, como a semântica de composição e semântica formal, mas que não serão abordadas aqui.

A semântica distribucional é uma área de pesquisa que se concentra no desenvolvimento de técnicas ou teorias que quantificam e categorizam semelhanças semânticas entre itens linguísticos com base em suas propriedades distributivas em grandes amostras de dados de texto. Vejamos um exemplo que lhe dê uma ideia do que quero dizer com semântica distribucional.

Suponha que você tenha dados de texto de blogs. Agora, você como pessoa sabe que macarrão, noodles, hambúrgueres e assim por diante são itens alimentícios comestíveis, enquanto que suco, chá, café e assim por diante são itens bebíveis. Como um ser humano, podemos facilmente classificar itens de alimentos bebíveis e comestíveis porque temos um certo contexto relacionado com cada um deles, mas as máquinas não podem realmente conhecer esse tipo de semântica. Há uma chance maior de que todos os itens alimentares descritos apareçam com certas palavras no conjunto de dados. Então, aqui estamos nos concentrando na distribuição de palavras em corpus e, digamos, que itens linguísticos ou palavras com distribuições similares têm significados semelhantes. Isso é chamado de hipótese distributiva.

Vejamos outro exemplo. Suponha que você tenha um conjunto de dados de documentos de pesquisa. Alguns dos trabalhos de pesquisa no conjunto de dados pertencem à categoria de engenharia e outros pertencem à categoria jurídica. Documentos com palavras como engenharia, equação, métodos e assim por diante estão relacionados à engenharia, portanto devem fazer parte de um grupo, e palavras como jurídico, advogado, institutos de advocacia e assim por diante estão relacionadas a documentos de pesquisa do departamento

jurídico, portanto, eles devem ser agrupados juntos. Usando técnicas de semântica distribucional, como **Word2vec**, podemos separar as diferentes palavras do domínio usando seus valores vetoriais. Todas as palavras com um significado similar são agrupadas porque têm uma distribuição similar no corpus. Você pode observar a figura abaixo, que mostra uma representação de um espaço vetorial de nosso dado exemplo de semântica distribucional, onde palavras contextuais semelhantes se juntam:



Word2vec é uma categoria de Semântica Distribucional: