CHRISTOFER FABIÁN CHÁVEZ CARAZAR

Resumen

Para poder ver la similitud semántica entre palabras, se utiliza el word-space model. Según Hinrich Schuze, el word-space se define como: Las parabras semanticamente relacionadas se ecunetran cerca, y las no relacionadas se encuentran distantes.
Geométricamente, el word-space model es una representación de significados de palabras, que puede ser representado en un espacio de n dimensiones, donde $n = \{1, 2, 3...\infty\}$. Nosotros, por nuestra capacidad física, no se nos hace posible visualizar espacios de gran dimensión. Un buen ejemplo para mostrar serían los espacions de 1 y dos dimensiones. Para una dimensión las palabras serían puntos en una recta, y para dos dimensiones las palabras serían puntos en un plano. En cualquiera de estos dos casos, la proximidad entre dos palabras indica cuan similar son sus significados.
Existen metáforas para obtener la similitud semántica. La similarity-is-proximity, que dice que más cerca esten las palabras, más similares son. La entities-are-locations, que dice que a fin de que dos palabras estan cerca, tienen que poseer la espacialidad; que necesitan para ocupar lugares en un espacio conceptual. Estas dos básicas metáforas constituyen la geometric metaphor of meaning, que dice que los significados entán ubicados en un espacio semántico, y la similitud emantica es la proximidad entre dos ubicaciones. El word-space model usa estadisticas sobre las propiedades distributivas de las palabras. De esta idea se genera la hipótesis distributiva, que dice que si dos palabras tienen propiedades distributivas, entonces tienen significados similares.

Para poder transformar las estadisticas a la geometría usaremos vectores de n dimensiones. Cuando tenemos un conjunto de palabras, generalmente en una oración, lo que tenmos que hacer es obtener el contexto de cada una de las palabras. Se forma una matriz cuadrada de la forma palabra$*$palabra y cada posición de la matriz toma una valor de 1 o 0 dependiendo si las palabras tienen relación o no.
Para hallar la similitud entre dos palabras, simplemente tenemos que calcular la inverse de la distancia entre los dos puntos. Otra forma de hallar la similitud es hallar la distancia Euclidiana entre los dos puntos.

Cuando implementamos un word space, nos encontramos con varios problemas, uno de ellos es manejar las dimensiones altas que haría que la matriz resultante sea demasiado enorme. Otro problema que encontramos es la escasez de datos, ya que la mayoría de los datos son 0s.
En el libro mensionan el algoritmo LSA que resuelve estos problemas. Este algoritmo utiliza una matriz words-by-documents, los datos son obtenidos por una fórmula indicada en el libro y utiliza el algoritmo SVD para reducir y reestructurar la dimencionalidad.
Otro algoritmo indicado en el libro es el HAL. Este algoritmo utiliza una matriz words-by-words, la distancia es calculada para los co-ocurrences, tiene concadenación row-column.