Vector Semantics

Dr.Furkan Göz

Aralık 4, 2023

Embeddings

- Kelimeleri temsil eden vektörler genellikle gömme (embeddings) olarak adlandırılır, çünkü kelime belirli bir vektör uzayına gömülür
- Benzer kelimeler anlamsal uzayda birbirine yakındır (Li ve arkadaşları (2015) duygu analizi için)

Embeddings

- Embeddings'ler anlam modelidir.
- Duygu analizinde:
  - Kelimelerle: aynı kelimenin eğitim ve testte olmasını gerektirir.
  - Embeddings'lerle: benzer kelimeler oluştuysa bu yeterlidir.
- Vector semantic models: pratiktir çünkü herhangi bir etiketleme olmadan metinden otomatik olarak öğrenilebilir.
- Örnek modeller:
  - Tf-idf modeli.
  - Word2vec modeli.

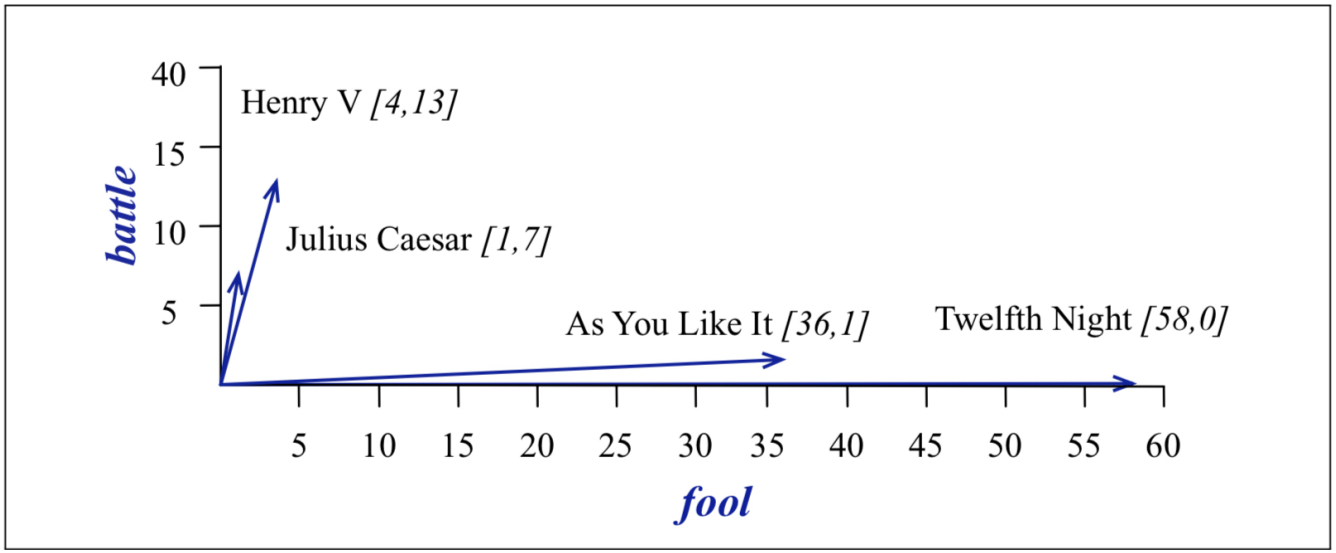
Term-Document Matrix

- Bir terim-belge matrisinde (1971, Salton), her satır sözlükteki bir kelimeyi temsil eder ve her sütun, belge koleksiyonundan bir belgeyi temsil eder.
- Shakespeare'in dört oyunundan dört kelimenin oluşumunu gösteren terim-belge matrisi:
- *battle* kelimesi, "Julius Caesar ve Henry V gibi tarih belgelerinde ortaya çıkan bir kelime türüdür", *fool* kelimesi ise "Twelfth Night gibi komedilerde ortaya çıkan bir kelime türüdür".

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Term-Docment Matrix

- *battle* (savaş) ve *fool* (aptal) kelimelerine karşılık gelen iki boyut
- Komedi türündeki oyunlar, *fool* boyutu için yüksek değerlere, *battle* boyutu için düşük değerlere sahiptir.



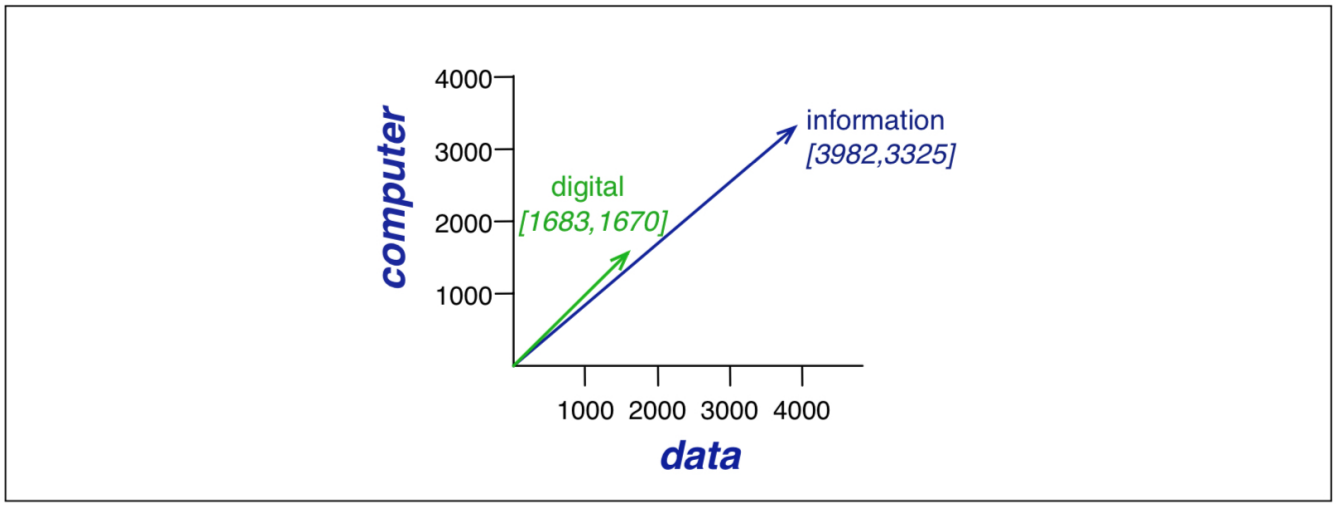
Word-word Matrix

- Term-term matrix ya da word-word matrix: Sütunlar belge yerine kelimelerle etiketlenir.
  - $|V| \times |V|$ boyutundadır ve her hücre, eğitim kümesinde bağlamlarda satır (hedef) kelime ile sütun (bağlam) kelimenin birlikte kaç kez geçtiğini gösterir.
- Kelime-kelime matrisindeki her satır, o satır (hedef) kelimesinin bağlam vektörünü oluşturur.
- İki kelimenin, bağlam vektörleri (co-occurence) benzerse anlamda benzerdir.

Word-word Matrix

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

- Bağlam vektörleri benzer olan iki sözcük anlam bakımından benzerdir



Cosinus Similarity

- İki hedef kelime v ve w arasındaki benzerliği tanımlamak için, iki vektörü alıp vektör benzerliğini hesaplayan bir ölçüt kullanılmalıdır.
- Bu ölçütlerden birisi, vektörlerin arasındaki açının kosinüsüdür.
- Kosinüs nokta çarpım operatörüne dayanır, bu operatör aynı zamanda iç çarpım olarak da adlandırılır.

$$\text{dot product}(\mathbf{v},\mathbf{w}) = \mathbf{v}\cdot\mathbf{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

Cosinus Similarity

- İki vektör *v* ve *w* arasındaki kosinüs benzerlik metriği şu şekilde hesaplanabilir:

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

- Kosinüs değeri, aynı yönü gösteren vektörler için 1; ortogonal vektörler için 0'dır.

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry}, \text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .018$$

$$\cos(\text{digital}, \text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

Örnek

Model, **information** kelimesinin **digital** kelimesine **cherry** kelimesinden daha benzer olduğunu göstermektedir.