Dil Modelleri

Mehmet Fatih AMASYALI

Kemik

İstatistiksel Dil Modelleri*

- Bir cümlenin olasılığını bulmak
- Kullanım alanları
  - Makine Çevirisi
    - P(taze balık aldım)>P(yeni balık aldım)
  - Yazım düzeltimi
    - Ali soan yedi
    - P(ali soğan yedi)>P(ali sokan yedi)
  - Konuşmadan Metne
    - P(soğan yedim)>P(sol an yedim)
  - vb.

https://web.stanford.edu/~jurafsky/slp3/slides/LM_4.pdf

Kemik

İstatistiksel Dil Modelleri

- Amaç 1: bir cümlenin olasılığını bulmak

  P(W)=P(w1,w2,…,wn)
- Amaç 2: bir kelime sekansından sonra gelecek kelimenin olasılığını bulmak

  P(w5|w1,w2,w3,w4)

  Bunları nasıl hesaplayacağız?

  Bayes Kuralı : P(A|B)=P(A,B)/P(B) → P(A,B)=P(B)*P(A|B)=P(A)*P(B|A)

  Zincir Kuralı: P(A,B,C,D)=P(A)*P(B|A)*P(C|A,B)*P(D|A,B,C)

Kemik

İstatistiksel Dil Modelleri

- Zincir kuralının genel hali:
- $P(w_1,w_2,w_3,\ldots,w_n)=$

  $P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\ldots P(w_n|w_1,\ldots,w_{n-1})$

  Örnek: P(ali,okuldan,buraya,geldi)= P(ali) * P(okuldan|ali) * P(buraya|ali,okuldan) * P(geldi|ali,okuldan,buraya)
- P(geldi|ali,okuldan,buraya) yı nasıl bulacağız?

  Bir olasılık :

$$P(\text{geldi}|\text{ali},\text{okuldan},\text{buraya}) = \frac{\text{frekans(ali okuldan buraya geldi)}}{\text{frekans(ali okuldan buraya)}}$$

Ama bu yolla güvenilir değerler elde etmek için yeterli büyüklükte bir derlemimiz elimizde olmaz. Çoğu frekans 0 olacaktır. Çünkü dil, diskrit uzayda çok seyrek. Peki ne yapalım?

Kemik

İstatistiksel Dil Modelleri

- Markov imdada yetişir ☺
- Markov varsayımı: her olasılık kendinden önceki k bileşene bağlıdır.
- k=0 için her olasılık bağımsızdır
  - P(w1,w2,w3,w4)=P(w1)*P(w2)*P(w3)*P(w4)
  - P(w4|w1,w2,w3)=P(w4)
  - P(geldi|ali,okuldan,buraya)=P(geldi)
- k=1 için her olasılık sadece bir öncekine bağlıdır
  - P(w1,w2,w3,w4)=P(w2|w1)*P(w3|w2)*P(w4|w3)
  - P(w4|w1,w2,w3)=P(w4|w3)
  - P(geldi|ali,okuldan,buraya)=P(geldi|buraya)

Kemik

- k=2 için her olasılık sadece bir ve iki öncekine bağlıdır
  - P(w1,w2,w3,w4)=P(w3|w1,w2)*P(w4|w2,w3)
  - P(w4|w1,w2,w3)=P(w4|w2,w3)
  - P(geldi|ali,okuldan,buraya)=P(geldi|okuldan,buraya)
- k değerini istediğimiz gibi arttırabiliriz.

Kemik

- Markov varsayımı ne yapıyor?
- Varsayım yokken
- $P(w_1,w_2,w_3,\ldots,w_n)=$
  $P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\ldots P(w_n|w_1,\ldots,w_{n-1})$
- $P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i|w_1, w_2, \ldots, w_{i-1})$
- Varsayım varken
- $P(w_i|w_1, w_2, \ldots, w_{i-1}) =$
  $P(w_i|w_1, w_2, \ldots, w_{i-k-1}, w_{i-k}, w_{i-k+1}, \ldots, w_{i-1}) \approx$
  $P(w_i|w_{i-k}, w_{i-k+1}, \ldots, w_{i-1})$

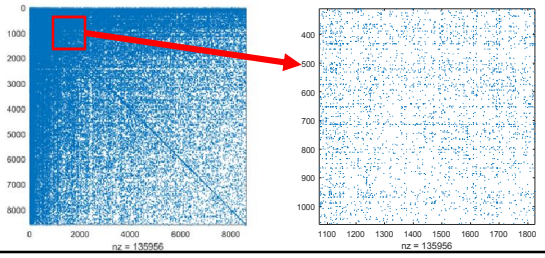
kırmızıları yok sayıyor

Kemik

- Markov varsayımının avantajları:
  - daha az işlem
  - daha az seyrek veriler
- Dezavantajı:
  - Dil, uzak bağımlılıklar içerdiğinden çok yeterli bir model değil
    - Çine yaptığım ziyarette … çok uğraştım ama bir türlü Çinceyi öğrenemedim.
    - Burada Çinceyi kelimesi belki onlarca kelime öncesindeki Çine kelimesine bağlı.
- K seçimi burada bir ikilem

Kemik

- 1000 ekonomi haberi
- 5 ten az geçen kelimeleri sildikten sonra yaklaşık 8K*8K lık bir matris
- Çok seyrek, ~64M den sadece ~140K 0 değil



Kemik

- ekonomi'den sonra en fazla geçen 5 kelime: $P(w_i|ekonomi)$ değeri en yüksek 5 kelime: bakanı, ve, ligi, bakanlığı, bankası
- para'dan sonra: politikası, cezası, ve, birimlerinin
- çok'dan sonra: önemli, daha, büyük, sayıda, iyi

Kemik

- Bir cümlenin bigram modeline göre olasılığı:
- P(fuar otomotiv sektörüne önemli bir hareket getirdi) =P(otomotiv|fuar) * P(sektörüne|otomotiv) * P(önemli|sektörüne) * P(bir|önemli) * P(hareket|bir) * P(getirdi|hareket)
- $P(w_i|w_{i-1}) = \frac{fr(w_{i-1},w_i)}{fr(w_{i-1})}$
- P(otomotiv|fuar)=fre(fuar otomotiv)/fre(fuar)
- Log(P(fuar otomotiv sektörüne önemli bir hareket getirdi)) = -21.2067

w_i	$fr(w_{i-1})$	$fr(w_{i-1},w_i)$	$P(w_i\|w_{i-1})$
fuar	18	1	1/18
otomotiv	78	4	4/78
sektörüne	35	1	1/35
önemli	560	162	162/560
bir	3940	4	4/3940
hareket	42	1	1/42

Kemik

0'larla başetmek Smoothing

- Log (P(fuar otomotiv sektörüne önemli bir hareket getirdi)) = log(P(otomotiv|fuar)) + log(P(sektörüne|otomotiv)) + log(P(önemli|sektörüne)) + log(P(bir|önemli)) + log(P(hareket|bir)) + log(P(getirdi|hareket)) = -21.2067
- Log (P(fuar otomotiv sektörüne önemli iki hareket getirdi)) = -INF
- **Add one estimation = Laplace smoothing**
- $P_{LS}(w_i|w_{i-1}) = \frac{fr(w_{i-1},w_i)+1}{fr(w_{i-1})+V}$
- Log (P_{LS}(fuar otomotiv sektörüne önemli bir hareket getirdi)) = -44
- Log (P_{LS}(fuar otomotiv sektörüne önemli iki hareket getirdi)) = -49

Kemik

En olası dizilişi bulmak

- Kelimeler = önemli, hareket, bir
- Olası dizilişler
- P_{LS}(önemli bir hareket)= -11.8570
- P_{LS}(önemli hareket bir)= -18.1881
- P_{LS}(bir önemli hareket)= -17.4616
- P_{LS}(bir hareket önemli)= -16.8926
- P_{LS}(hareket önemli bir)= -13.0944
- P_{LS}(hareket bir önemli)= -17.4034
- Smoothing kullanmasaydık?
- Bir başka uygulama: Şu kelimelerle bir cümle kurun

Kemik

Dil modellerini değerlendirmek

- Elimizde 2 dil modeli olsun (A, B). Hangisi daha iyi?
- Harici yöntem: her 2 modeli de farklı görevler için (Makine Çevirisi, Yazım düzeltimi, Konuşmadan Metne) kullan. O görevlerdeki performanslarına göre karşılaştır
- Dahili yöntem: Derlemi eğitim ve test kümesi olarak ayır. A ve B yi eğitim üzerinden oluştur.
  - Test kümesindeki cümlelere hangisi daha yüksek olasılık veriyorsa o daha iyidir.
  - Bir cümlenin başı verildiğinde sonunu hangisi doğru tahmin ediyorsa o daha iyidir.
  - Bu görevlerde Markov varsayımında k=0 seçimi nasıldır?

Kemik

Dil modelleriyle cümle / dizilim üretimi

- Bir ilk kelime seçelim. Bundan sonra gelecek kelimeleri dil modeli ile belirleyelim.
- K=0 için her zaman en yüksek olasılığı seçersek hep aynı kelime tekrar eder
- K=1 için, her zaman en yüksek olasılığı seçersek bir kelimeden sonra hep aynı kelime gelir.
- K=2 için, ardışık 2 kelimeden sonra hep aynı 3. kelime gelir.
- Bunu aşmanın 2 yolu var
  - En yüksek olasılıklı t taneden birini rasgele seçmek
  - Hepsi içinden olasılıklarına göre seçim yapmak (Shannon Game)
- Ne zaman duracağız: İstediğimiz sayıda kelime / cümle ürettiğimizde

Kemik

- Bigram (K=1) modeli ile
- En yüksek olasılıklı t taneden birini rasgele seçer. t değeri azaldıkça metinlerde kopya çekme olasılığı artar. t=5
- 1000 ekonomi haberi ile
  - "fonların yüzde 5 de bu konuda son 10 ın da çok sayıda türk telekom gibi çok önemli olduğunu belirten bakan şimşek 2012 nin."
  - "mesleki eğitim ve ticaret bakanlığı ın piyasa ile ilgili olarak belirlendi bu nedenle bir önceki yıla da en fazla artış oranı ise lira oldu bu."
  - "akademik araştırmalar sonucunda yer aldığı bilgiye göre bir şekilde diyen ve yüzde 10 yılda bir süre içerisinde türkiye istatistik verilerine borsa seçimlerinin dikkati çekti türkiye."
- Burada first order (K=1) yapının problemi bariz görünüyor. "yer aldığı bilgiye" "yüzde 10 yılda"

Kemik

- trigram (K=2) modeli, t=5
- 1000 ekonomi haberiyle üretilen cümleler
  - "yaşanan sıkıntıları için devamlı ucuz bilet satıyoruz 100 90 ı aslında biz bu bitireceğiz inşallah yıl sonuna kadar türkiyede üretilen fazlasını belirterek geçen yıl kasım ayı ihracatını"
  - "birlikte fiyatlardaki üzerindeki etkileri takip edilecek gerek halinde düzenlemede birtakım yeni getirilecek talepler daha önceki bir açıklamasında 2012 yılının ilk 2 ihaleyi 500 adet olacak satış faaliyet"
- 1000 magazin haberiyle üretilen cümleler:
  - "arkadaşlarının da kendisini hiç söyleyen nilüfer kanser olduğunu tüm anlattı ilk itibaren dizinin kerem benim olduğu gibi bu hafta muhteşem neler olacak ali ikna eder ama şimdi"
  - "oyuncuların eğitimi yoktu ama hep başka şeyleri yapmak için burada bu arkadaşlar benim dönüş ve benim gibi tabii ki oldu ama devam çok bir adam alman sevgilisiyle"
- Ardışık 3 lüler daha anlamlı artık.

Kemik