YZM 511 - İstatistiksel Yapay Öğrenme

Doğal Dil İşlemeye Giriş I

11/12/2024

Doğal Dil İşleme (NLP) Nedir?

- İnsan dillerini (doğal diller) anlamayı, işlemeyi ve oluşturmayı amaçlayan bir yapay zeka (AI) alanıdır.

- Dilbilim, bilgisayar bilimi ve yapay zeka tekniklerini birleştirir.

- Bilgisayarların insan dilini: anlamasını, analiz etmesini, anlamlı bir şekilde yanıt vermesini sağlar.

NLP'nin Ana Amaçları

- **Dil Anlama (Natural Language Understanding - NLU)**
İnsan dilinin anlamını çözümlemek ve bağlamını anlamak.

- **Dil Üretimi (Natural Language Generation - NLG)**
İnsan benzeri doğal bir dilde metin veya konuşma üretmek.

NLP'nin Kullanım Alanları

- **Makine Çevirisi**
  Örneğin, Google Translate veya DeepL gibi araçlarla diller arası çeviri.
- **Sesli Asistanlar**
  Siri, Alexa, Google Asistan gibi asistanların doğal dille konuşabilmesi.
- **Metin Madenciliği ve Anlam Çıkartma**
  Büyük metin yığınlarından önemli bilgilerin çıkarılması.
- **Duygu Analizi**
  Sosyal medya, incelemeler veya müşteri geri bildirimlerindeki duygu durumunu analiz etmek.

NLP'nin Kullanım Alanları

- **Sohbet Botları**
  Şirketler için müşteri hizmetleri sunan botlar.
- **Arama Motorları**
  Google ve Bing gibi motorların doğru sonuçlar göstermesi.
- **Otomatik Özetleme**
  Uzun metinlerin kısa ve anlamlı özetlerini oluşturmak.
- **Dil Modelleme**
  GPT gibi modellerin geliştirilmesi.

Bankaya gittim ve … çektim.

Metin Ön İşleme

NLP'deki her analiz ve modelleme süreci, düzgün hazırlanmış bir metin verisiyle başlar. Metin, doğal dilde olduğu için genelde gürültülü ve hamdır. Çünkü:

- Dilbilgisi kurallarına uymayan ifadeler içerir.
- Gereksiz kelime ve semboller bulunabilir.
- Eksik veya bağlama katkı sağlamayan ifadeler olabilir.

Düzensiz Dilbilgisi

Doğal dilde insanlar, dilbilgisi kurallarına her zaman tam olarak uymayabilir.

Ham Metin:
"Bugün hava çok güzel gidiyo ama bilmiyom dışarı çıkayım mı? "

Problemler:

- "gidiyo" ve "bilmiyom" kelimeleri yazım hatalı (düzensiz yazılmış).
- Dilbilgisi eksiklikleri var: "dışarı çıkayım mı?" bağlamsal olarak tam değil.

Çözüm: Bu tür ifadeler normalleştirilmeli:
"Bugün hava çok güzel gidiyor ama bilmiyorum dışarı çıkmalı mıyım?"

Fazla Bilgi veya Alakasız Kelimeler

Metinlerde gereksiz kelimeler, tekrarlamalar veya bağlama katkısı olmayan ifadeler bulunabilir.

Ham Metin:
"Evet yaaa çok çok güzel ya, bence bence harikaaaa bir şey yaaa!!! «

Problemler:

- "yaaa", "çok çok", "harikaaaa" gibi ifadeler bağlama katkı sağlamıyor.
- Aynı kelimeler tekrarlanıyor ("çok", "bence").

Çözüm: Bu tür tekrarlar ve gereksiz ifadeler temizlenebilir:
"Evet, bence harika bir şey."

Özel Karakterler ve Noktalama İşaretleri

Ham metinlerde, analiz için gereksiz olan özel karakterler, semboller veya fazla noktalama işaretleri bulunabilir.

Ham Metin:

"Selam!!! 😊 😊 😊 Bugün hava nasıl??? 😎 🌞 «

Problemler:

- Fazla sayıda noktalama işareti ("!!!", "???")
- Emojiler (" 😊 ", " 😎 ")

Çözüm: Noktalama işaretleri ve emojiler temizlenebilir:

"Selam. Bugün hava nasıl?"

Durdurma Kelimeleri (Stopwords)

Bazı kelimeler anlam taşımadığı ya da modele bilgi katmadığı için "gürültü" sayılır.

Ham Metin:
"Ve sonra, o zaman, işte aslında o olay çok ilginçti."

Problemler:

- "Ve", "sonra", "o zaman", "aslında" gibi kelimeler cümlede bilgiye katkıda bulunmuyor.

Çözüm: Bu kelimeler çıkarılarak metin sadeleştirilir:
"O olay çok ilginçti."

Ham Metin:

"Bugün hava müthiş yaaaa!!! Şey, hani şu dışarı çıkalım mı dedik ya, o muhteşem olurmuş yaa bence. Hani ben şey düşündüm, eğer çıkarsak ne yapsak ki? Belki yürüyüş falan güzel olabilir mi? 🤔 😊 "

Düzenlenmiş Metin:

"Bugün hava çok güzel. Dışarı çıkmayı düşündüm. Belki yürüyüş yapmak iyi bir fikir olabilir."

Tokenizasyon (Metni Parçalama)

Metni kelime veya cümle düzeyinde küçük parçalara bölmek (**tokenizasyon**) doğal dil işleme (NLP) projelerinin temel bir adımıdır.

"Bugün hava çok güzel."

["Bugün", "hava", "çok", "güzel"]

Lemmatizasyon veya Kök Bulma

Kelimenin sözlükteki temel formunu bulmak anlamına gelir.

"koşuyor", "koştu" → "koş "

"evden", "eve", "evler" → "ev“

"büyükçe", "büyüktü", "büyüyordu" → "büyük"

OZET

1. Standardize Etme

Tüm harfleri **küçük** veya **büyük harfe** çevir.

Örnek: "Bugün Hava Harika!" → "bugün hava harika"

2. Noktalama İşaretleri ve Emojileri Kaldırma

Anlamsal katkı sağlamayan **noktalama işaretlerini** ve **emojileri** temizle.

Örnek: "Bugün hava harika! 😊 " → "bugün hava harika"

3. Stopwords'leri Kaldırma

"ve", "çok", "mi", "ama" gibi bağlama katkısı olmayan kelimeleri çıkar.

Örnek: "Bugün hava çok güzel." → "bugün hava güzel"

4. Tokenize Etme

Metni **kelimelere** veya **cümlelere** böl.

Örnek: "Bugün hava güzel." → ["bugün", "hava", "güzel"]

5. Köklerini Ayırma

Kelimeleri temel köklerine indir.

Örnek: ["koşuyorum", "koştu", "koşacak"] → ["koş", "koş", "koş"]