Text Classification and Naive Bayes

Metin Sınıflandırma ve Naïve Bayes

Metin Sınıflandırılması

- Naive Bayes sınıflandırıcı temel bir metin sınıflandırıcıdır.
- Bu sınıflandırıcı, metin sınıflandırmada pek çok sürecin temelini oluşturur.

Metin Sınıflandırma ile Çözülebilecek Problemler: Aşağıdaki bir spam midir?

Good morning Dan,

Please familiarize yourself with the attached file.
Reply here if you have any questions.

Thank you.

John and Mike,

Appreciate your flexibility this week, as the team navigates the sensitivities surrounding some of the project work taking place at the sites. Please tentatively plan for mobilization on 05/16/2022, in order to begin the final stages of the upgrade.

I will follow-up tomorrow with a confirmation if all indications are we will be given the “all-clear” before EOB Wednesday/SOB Thursday.

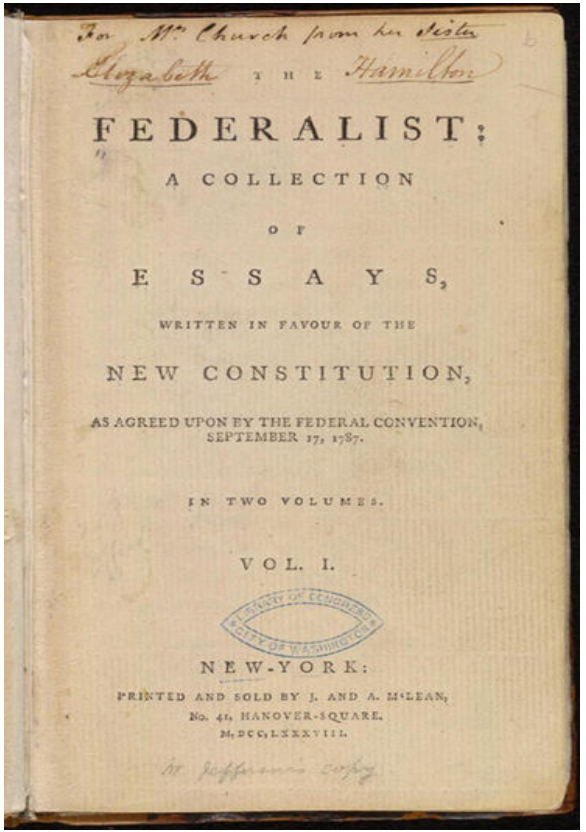
Appreciate your support.

Regards,

Judy Sewell
Project Manager

Metin Sınıflandırma ile Çözülebilecek Problemler: Federalist Papers yazarları kimdir?

- 1787-1788 yıllarında isimsiz olarak (anonymously) yazılmış makalelerdir.
  - New York halkını, ABD Anayasası'nı onaylamaya ikna etmeya çalışan makalenin yazarları bilinmemektedir.
- Bu makaleler aslında Alexander Hamilton, James Madison ve John Jay tarafından yazılmıştır.
  - 1963 yılında Mosteller ve David L. Wallace isimli araştırmacıları Bayes metotlarını kullanarak makalelerin yazarlarını çözmüşlerdir.



Metin Sınıflandırma ile Çözülebilecek Problemler: Pozitif ya da Negatif film eleştirilerinin sınıflandırılması


İnanılmaz derecede hayal kırıklığına uğrattı.


Çılgın karakterlerle dolu, zengin bir şekilde uygulanmış hiciv ve bazı harika olay örgüsü dönüşleri gördük.


Bu şimdiye kadar çekilmiş en iyi acayip komedi acıklıydı.


En kötü yanı boks sahneleriydi.

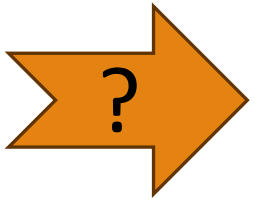
cümlelerin tipleri olumlu / olumsuz olarak sınıflandırılabilir.

Metin Sınıflandırma ile Çözülebilecek Problemler: Makalenin konusu nedir?

MEDLINE Article

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Rosanna Gahl,* Lisa Menn,* Gail Ramsberger,* Daniel S. Jurafsky,* Elizabeth Elder,* Molly Rewega,* and L. Holland Audrey*



MeSH Subject Category Hierarchy*

Antogonists and Inhibitors

Blood Supply

Chemistry

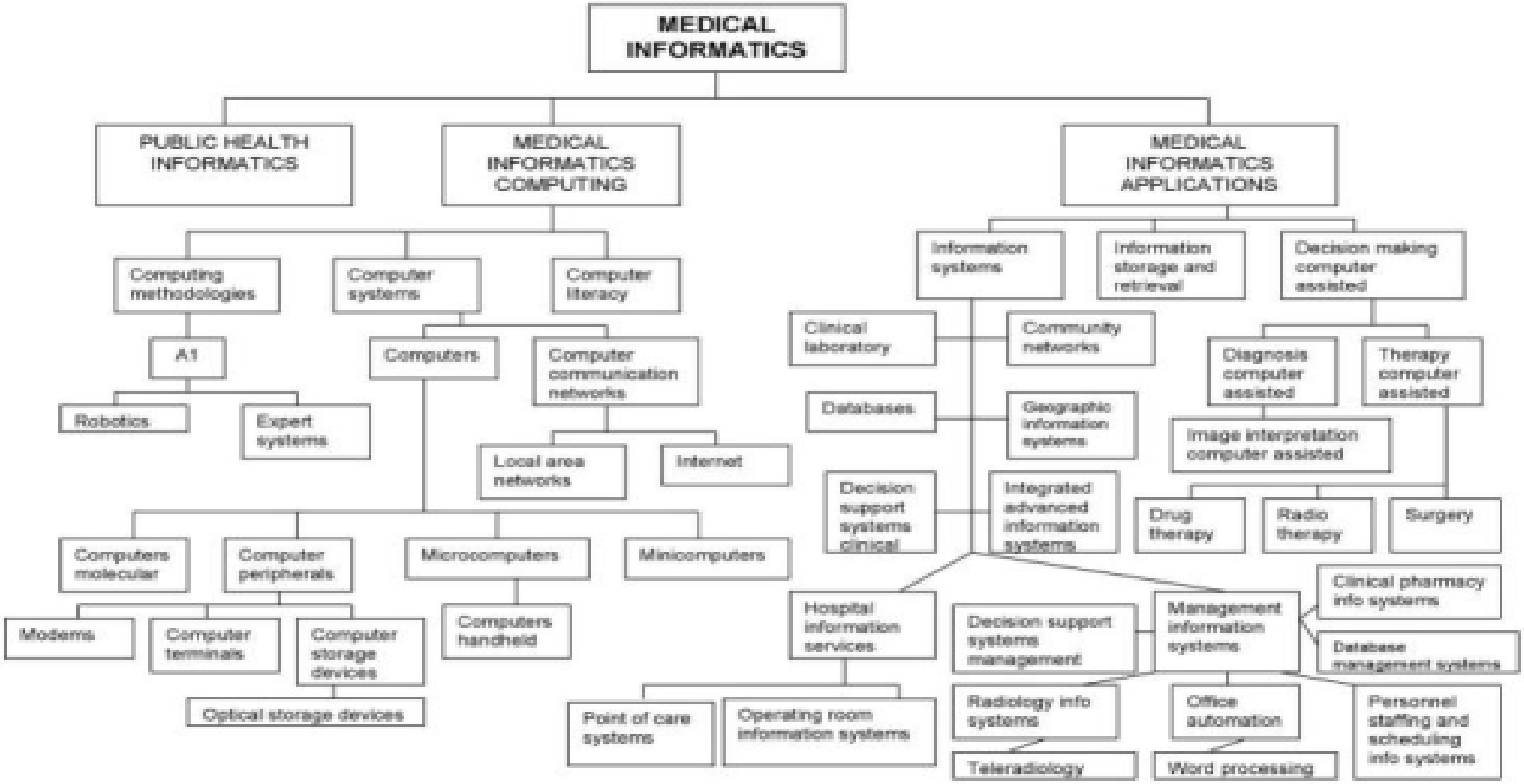
Drug Therapy

Embryology

Epidemiology

*Medical Subject Headings (MeSH)

MeSH Hiyerarşik İlişkilerine İlişkin Örnek



Metin Sınıflandırma ile Özet Olarak:

- Konuların kategorilerine ayrılması, makale başlıkları veya makale türleri ile ilgili farklı atamalar yapılabilir.
- Spam tespiti yapılabilir.
- Yazar tanımlama yapabilir.
- Makalenin hangi dilde yazıldığına ilişkin tahmin yapılabilr.
- Duygu analizi yapılabilir
- ……

Metin Sınıflandırma Tanımı

Giriş:

Bir d dokümanı ve

Sabit bir sınıflar dizisi $C = \{c_1, c_2, \ldots, c_J\}$

Çıktı: Tahmin edilen herhangi bir sınıf $c \in C$

Sınıflandırmanın amacı herhangi bir gözlemin ele alınması, bundan bazı özelliklerin çıkarılması ve buradan da ilgili gözlemi farklı sınıflardan birinde sınıflandırmaktır.

En Temel Sınıflandırma Metodu: Elle Oluşturulmuş Kurallar

- Metin sınıflandırma yöntemlerinden biri, insanların elle yazdığı kuralları kullanmaktır.
- Elle oluşturulmuş kural tabanlı sınıflandırıcılar, dil işlemedeki son teknoloji sistemlerinin bileşenlerinden biri olabilir.
  - Bu kuralların kırılgan olma olasılığı yüksektir. Çünkü;
    - ✓ Durumlar ya da veriler zamanla değişir ve bazı süreçlerde insanlar kuralları oluşturmada iyi olmayabilir.
- Kurallar kelime kombinasyonlarına göre oluşturulabilir.
  - Spam Örneği : Kara liste adresi VEYA ("dolar" VE "seçildi /has been chosen")
- Özel alanlarda kurallar uzmanlar tarafından dikkatlice düzenlenirse doğruluk yüksek olabilir. Ancak:
  - Kuralları oluşturmak ve sürdürmek pahalıdır
  - Özeldirler; yüksek hassasiyet içerebilirler.

Denetimli Sınıflandırma Metodu/ Supervised Machine Learning

Girdi:

- *Bir d dokümanı*
- Sabit sınıflardan oluşan bir dizi $C = \{c_1, c_2, \ldots, c_J\}$
- *m tane elle oluşturulmuş eğitim kümesi* $(d_1,c_1),\ldots,(d_m,c_m)$

Çıktı:

- Öğrenmiş bir sınıflandırıcı $y:d \rightarrow c$

Denetimli Sınıflandırma Metotları

- Pek çok sınıflandırıcı metot vardır. Bunlardan bazıları:
  - Naïve Bayes
  - Logistic regression
  - Neural networks
  - *k*-nearest neighbors
  - ….
- Aynı zamanda ön eğitimli dil modelleri kullanılabilir
  - Sınıflandırıcılar olarak ince ayar (Fine-tuned)
    - ✓ Önceden eğitilmiş bir dil modelini belirli bir göreve /alana özgü verilerle yeniden eğitme sürecidir.
    - ✓ Daha sonra bir sınıflandırma yapılması istenir.

Duygu Analizi /Sentiment Analysis

- Duygu analizi, duygunun dış analizi, bir yazarın bir nesneye karşı ifade ettiği olumlu veya olumsuz yönelimleri olabilir.
- Duygu analizinin en basit versiyonu ikili sınıflandırma yapılmasıdır.
- İncelenen kelimelerle doğru ipuçları belirlenir.
  - Örneğin, film ve restoranların olumlu ve olumsuz incelemelerinden alınan bazı sözcükler şunlar olsun: «great, richly, awesome, pathetic, awful, ridiculously» gibi kelimeler bilgilendirici ipuçlarıdır.

    \+ .....zany characters and richly applied satire, and some great plot twists

    \- It was pathetic. The worst part about it was the boxing scenes...

    \+ .....awesome caramel sauce and sweet toasty almonds. I love this place!

    \- .....awful pizza and ridiculously overpriced...

Spam Tespiti / Detection

- Spam tespiti, bir e-postayı spam veya spam olmayan şeklinde iki sınıftan birine atama görevi gerçekleştirir.
  - Bu da ikili sınıflandırmadır.
- Bu sınıflandırmayı gerçekleştirmek için birçok sözcüksel özellik ya da farklı özellikler kullanılabilir.

Metin Sınıflandırılması ve Naive Bayes

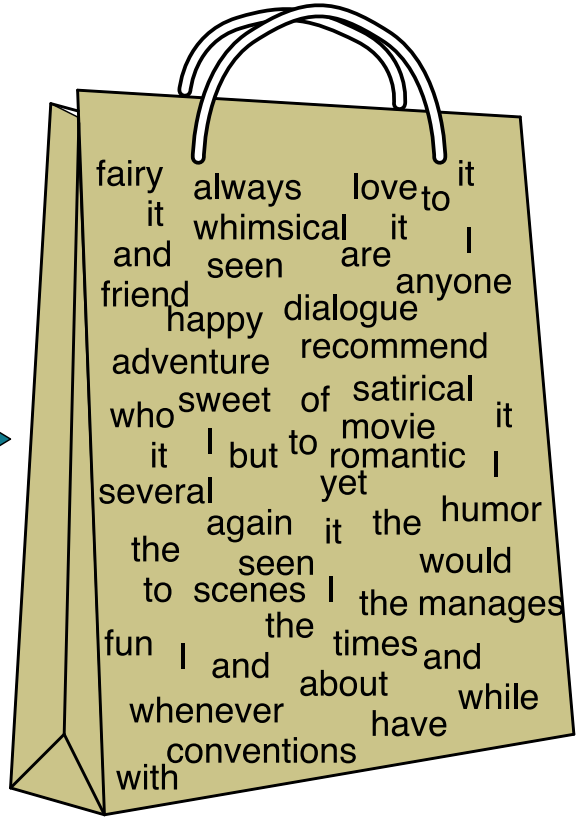
Naive Bayes Classifier

Naive Bayes Sezgisi

- Basit ("naive") sınıflandırma yöntemi Bayes kuralına dayanır.
  - Sezgi, bir dokümanın basit olarak betimlenmesine dayanır.
- Bir metin belgesi, bir kelime torbasıymış gibi, yani konumları göz ardı edilerek sıralanmamış bir kelime kümesiymiş gibi temsil edilir.
  - Belgede geçme sayıları / frekansları korunur.

The Bag of Words /Kelime Torbası Betimlemesi

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Kelime Torbası Betimlemesi

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

$\gamma($ $)=c$

- Sezgisel olarak ifadesi kelime torbası varsayımıdır.
- Konumun, sözcüğün geçtiği yerin önemli olmadığı, herhangi bir kelimenin belgedeki 1., 20. veya son kelime olarak geçmesinin sınıflandırma üzerinde aynı etkiye sahip olduğunu varsayılır.

Doküman ve Sınıflara uygulanan Bayes' Kuralı

Bir d dokümanı ve bir *c sınıfı için:*

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naive Bayes Sınıflandırıcı (I)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP "maximum posteriori /deneysel " = en olası sınıf (most likely class)

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Payda elimine edilir.

Naive Bayes Sınıflandırıcı (II)

"Likelihood/ olası "

"Prior/önceki "

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(f_1, f_2, ..., f_n|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

Dokümanının $f_1..f_n$ özellikleriyle betimlenmesi

O(|X|^n•|C|) parametrelidir

Çok fazla sayıda eğitim örneği mevcutsa tahmin edilebilir.

Bu sınıf ne sıklıkla gerçekleşir?

Bir derlemdeki görece /relatif frekanslar sayılabilir.

Çok terimli Naive Bayes Bağımsızlık Varsayımları

Bag of Words /Kelime Torbası Varsayımı : Pozisyonun önemi olmadığı kabul edilmektedir.

Şartlı Bağımsızlık: Bir c sınıfı verildiğinde özellik olasılıkları $P(f_i|c_j)$ bağımsız (independent) olasılıklardır

$$P(f_1, f_2, ..., f_n|c) = P(f_1|c) \cdot P(f_2|c) \cdot ... \cdot P(f_n|c)$$

Çok terimli / multinomial Naive Bayes Sınıflandırıcı

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{f \in F} P(f|c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in positions} P(w_i|c)$$

Pozisyon ??? Test dokümanındaki tüm sözcük pozisyonlarıdır

Fazla Sayıda Olasılığın Hesaplanması Problemi

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in positions} P(w_i|c)$$

Pek çok olasılığın çarpımı yapılarak sonuç elde edilmektedir. Örneğin;

.0006 * .0007 * .0009 * .01 * .5 * .000008….

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} \log P(c) + \sum_{i \in positions} \log P(w_i|c)$$

log uzamı (space) ile fazla sayıda hesaplama yapabilmek mümkündür.

Logaritmik özellik kullanılarak, $\log(ab) = \log(a) + \log(b)$, olasılıkların çarpılması yerine olasılıkların logaritmaları toplanacaktır.

Özetle:

- Logaritma almak sınıfların sıralamasını değiştirmez.
  - En yüksek olasılığa sahip sınıf aynı zamanda en yüksek logaritmik olasılığa sahiptir
- Bu doğrusal bir modeldir:
  - Sadece ağırlıkların toplamının maksimumu alınır.
    - ✓ Bu, girdilerin doğrusal bir fonksiyonudur.
  - Sonuç olarak: Naive Bayes doğrusal (lineer) bir sınıflandırıcıdır.

Metin Sınıflandırma ve Naïve Bayes: Öğrenme

The Naive Bayes Classifier

Naive Bayes: Learning

Çok Terimli (Multinomial) Naive Bayes Model P(c) olasılığı nasıl tahmin edilir?

Maximum olasılıklı tahminler (max. likelihood estimates) için veri setindeki sözcüklerin frekansları kullanılır.

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

Eğitim verisinin c sınıfındaki belge sayısı N_c , toplam belge sayısı N_{total} olsun.

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

c kategorisindeki tüm belgeler tek bir "kategori c" metninde birleştirilir. Bu birleştirilmiş belgedeki w_i sıklığı kullanılarak olasılığın maksimum olabilirlik tahmini verilir.

Çok Terimli (Multinomial) Naive Bayes Model P(fi |c) olasılıkları nasıl öğrenilebilir?

P(fi |c) olasılığını öğrenmek için, herhangi bir özelliğin yalnızca belgenin kelime torbasından bir kelimenin olduğu varsayılır.

Bu nedenle, c konusunu içeren tüm belgelerdeki kelimelerin tümü arasında w_i kelimesinin görünmesi sayısı olarak P(wi |c) hesaplanır.

Parametre Tahmini

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

w_i sözcüğü, tüm sözcükler arasında c_j konusunu içeren dokümandaki tüm sözcükler arasında bulunur.

V sözlüğü, tüm sınıflardaki sözcük tiplerinin tümünün birleşimidir; sadece bir c sınıfında bulunan sözcükler değildir.

- Önce herhangi bir j konusunda bir mega doküman oluşturulur.
  - Bu mega doküman, ilgili konudaki tüm dokümanların art arda eklenmesi ile oluşturulur.
  - Mega dokümandaki w frekansı kullanılır.

Maximum Likelihood ile Eğitimde Problem

Örneğin; *fantastic* kelimesinin geçtiği ve konu başlığının pozitif olarak sınıflandırıldığı hiçbir eğitim dokümanı olmasın.

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

Zero probabilities, diğer koşullar ne olursa olsun koşul oluşturmayacaktır.

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in positions} P(w_i|c)$$

Naive Bayes ile tüm olasılıklar basit olarak çarpıldığı için, herhangi bir sınıf için sıfır olasılık, diğer sonuçlar ne olursa olsun, sınıfın olasılığının sıfır yapar.

Naive Bayes için Laplace (add-1) Smoothing
Laplace Düzeltmesi

- Toplam olasığın 0 olması durumunu kaldırmak için farklı bir yöntem önerilir.
  - Bu algoritmaya add-one (Laplace) smoothing / Laplace düzeltmesi adı verilir.
- Olasılıkları normalleştirmeden önce tüm n-gram sayımlarına bir eklenir.
  - Daha önce sıfır olan sayımların sayısı a 1 olacak, 1'li sayılar 2 olacaktır.

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right)+|V|}$$

Naive Bayes için Laplace add-1 Smoothing Laplace Düzeltmesi

Kelime dağarcığında V adet kelime bulunur.

Her bir (w,c) için bir arttırım yapıldığı için, payda da ekstra V adet gözlem yapılacak şekilde ayarlama yapmak gerekir. .

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right)+|V|}$$

Bilinmeyen Sözcükler

- Bilinmeyen sözcükler nelerdir?
  - Test datasında görünen ana eğitim datası ya da sözlükte olmayanlardır
- Bu sözcükler ihmal edilir.
  - Test dokümanından kaldırılır.
  - Test dokümanında olmadıkları varsayılır
  - Bu sözcükler hiç bir olasılığa dahil edilmez
- Niçin bilinmeyen bir sözcük modeli oluşturulmaz?
  - Hangi sınıfta daha fazla bilinmeyen kelime olduğunun bilinmesi genellikle yararlı değildir.

Multinomial Naïve Bayes: Learning

Eğitim datasından sözlük (Vocabulary) oluşturulur.

Calculate $P(c_j)$ terms

- For each c_j in C do

  $docs_j \leftarrow$ all docs with class $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Stop words

- Bazı sistemler durdurma sözcüklerini yok sayar.
  - Durdurma sözcükleri: «the» ve «a» gibi sık kullanılan sözcüklerdir.
  - Eğitim setindeki sözcük sıklığına göre kelime dağarcığını sıralar.
    - ✓ En iyi 10 veya 50 sözcüğü durdurma sözcüğü listesi olarak adlandırılır.
  - Hem eğitim hem de test setlerinden tüm durdurma sözcüklerini kaldırılır.
- Ancak durdurma sözcüklerini kaldırmak genellikle işe yaramaz
  - Bu yüzden pratikte çoğu NB algoritması tüm sözcükleri kullanır ve durdurma sözcüğü listelerini kullanmaz

Metin Sınıflandırma ve Naive Bayes

Naive Bayes: Learning

Sentiment ve Binary Naive Bayes

Sentiment Örneği

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

Sentiment Örneği : 1 smoothing /düzeltme

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable ~~with~~ no fun

1. Prior from training:

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

P(-) = 3/5
P(+) = 2/5

2. Drop "with"

3. Eğitimden en olasılar /likelihoods

$$p(w_i|c) = \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c)\right) + |V|}$$

$$P(\text{“predictable”}|-) = \frac{1+1}{14+20} \qquad P(\text{“predictable”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“no”}|-) = \frac{1+1}{14+20} \qquad P(\text{“no”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“fun”}|-) = \frac{0+1}{14+20} \qquad P(\text{“fun”}|+) = \frac{1+1}{9+20}$$

4. Test kümesini puanlama :

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$