# Web Document Classification

Web pages, Text extraction, The Perceptron, PyTorch

## Web Search

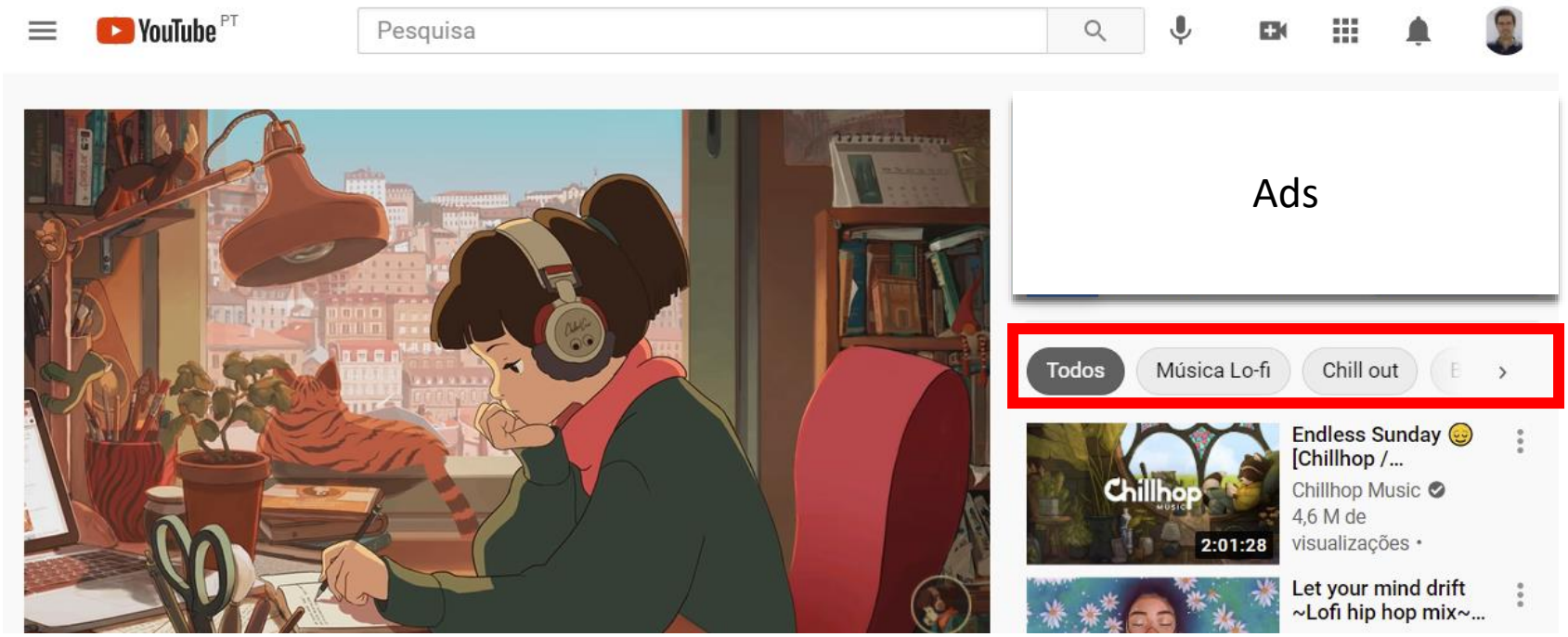
# Information classification on the Web

- There are many classification problems on the Web
- Data is highly heterogeneous
- Tasks are also very heterogeneous
- There's never one solution for all problems
  - a.k.a. there is no silver bullet in CS

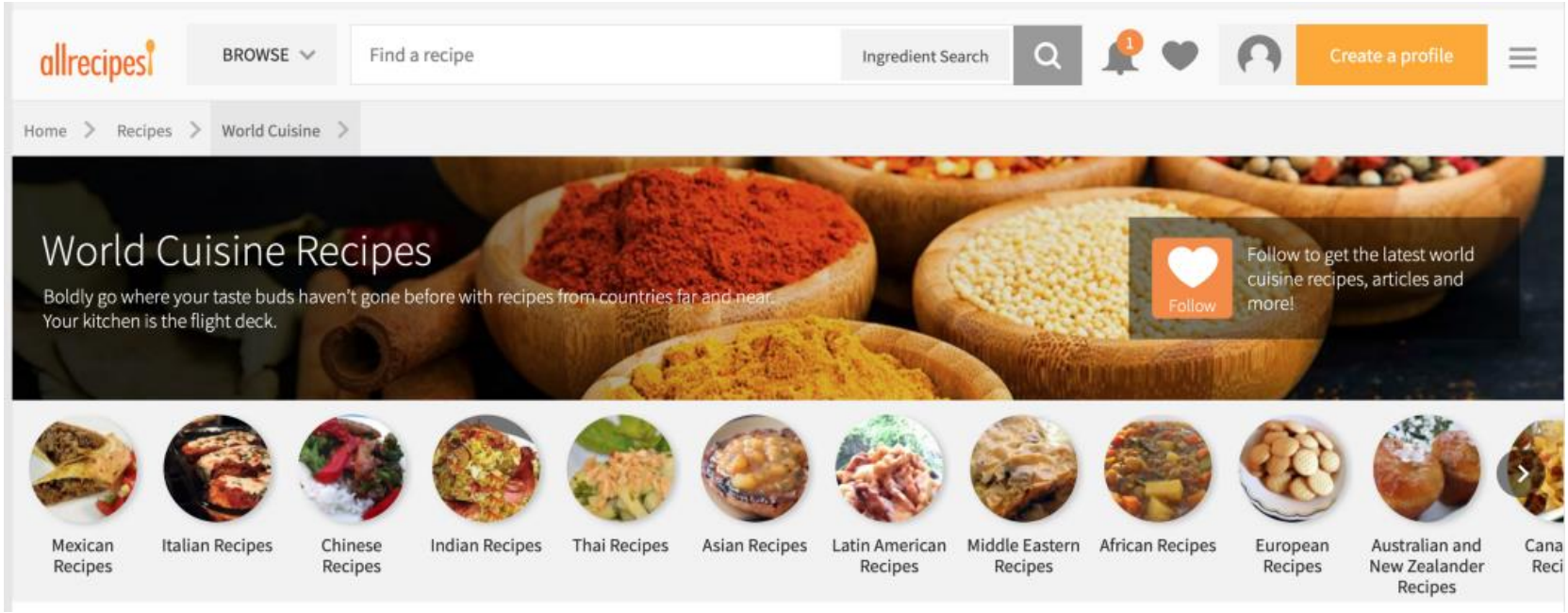
# News categories



# YouTube video tags



# Recipes


allrecipes
BROWSE
Find a recipe
Ingredient Search
Create a profile
Home
Recipes
World Cuisine
World Cuisine Recipes
Boldly go where your taste buds haven't gone before with recipes from countries far and near.
Your kitchen is the flight deck.
Follow
Follow to get the latest world cuisine recipes, articles and more!
Mexican Recipes
Italian Recipes
Chinese Recipes
Indian Recipes
Thai Recipes
Asian Recipes
Latin American Recipes
Middle Eastern Recipes
African Recipes
European Recipes
Australian and New Zealander Recipes

# Rating prediction and review spam


Proscenic M6 PRO Wi-Fi Connected Robot Vacuum Cleaner and Mop, Alexa & Google Home & App Control, Lidar Navigation, Robotic Vacuum with Mapping, 2600 Pa Suction and Selective Room Cleaning

Visit the Proscenic Store

4.4 out of 5 stars 1,835 ratings

Price: $369.00 + $144.76 Shipping & Import Fees Deposit to Portugal Details

| | |
|---|---|
| **Brand** | Proscenic |
| **Color** | Gray |
| **Surface Recommendation** | Carpet |
| **Controller Type** | Amazon Alexa, Voice Control |
| **Battery Cell Composition** | Lithium |

**About this item**

- 【Lidar navigation】L and maps your home

Mika

★★★★★ **Almost gave up due to WIFI, glad I didn't**

Reviewed in the United States on January 15, 2021

**Verified Purchase**

The app is so much easier to use (once synced) than the Roborock. It cleans edges like a beast, it's not cranky noisy, it doesn't bump into things as hard as the roborock did. It mapped out my whole downstairs perfectly.

Here's why I almost gave up. The app kept crashing when I tried to sync it with my iPhone, over and over and I was hurt cause I liked the way it cleaned and the app is 10X more handy than the remote. Then I remembered how easy our Elliptical/iFit synced with hubby's droid but wouldn't with my iPhone and said hmmmm, lemme download the app on his phone and it all set up. Sure enough. It synced the FIRST time!!! I saw that my whole downstairs has already been mapped out by the vacuum, I just had to draw lines to separate the rooms and name them. What a relief. I hate to spend this much money on something and not LOVE it completely ya know. I logged out on his phone, logged back in on my iPhone and sure enough all the data and settings were there. I had even changed my WIFI settings thinking it was that! OMG. Nothing is perfect but I think this one has some great features, it seems quite meticulous under the kitchen table a s chairs like it's determined to pick up every speck-a-dust! For the features and price point, I'm happy so far.

6 people found this helpful

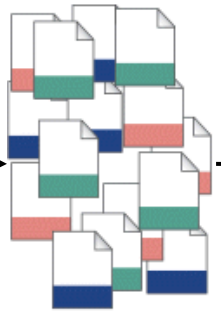
# Challenges

- Consider a set of HTML documents:
  - How to extract content (text and video) from structure (HTML tags)?
  - How to represent content?
  - Which categories?
  - How to classify documents?

# Web Document Categorization

**HTML documents**


**Uncategorized text documents**



**Taxonomy**

Economy

Recipes

Sports

...

14 ABR 2021
Nº 55511


Índice



# Atraso com vacina da Johnson não altera metas de plano português

Task-force para Plano de Vacinação contra a Covid-19 diz que meta de ter 70% da população vacinada até ao fim do verão está de pé, apesar de a vacina da Johnson & Johnson estar agora suspensa nos EUA e ir chegar mais tarde à Europa.


Desde o início do processo de vacinação, já foram administradas mais de 6,8 milhões de doses da vacina da Johnson & Johnson nos EUA.

Ana Mafalda Inácio
13 Abril 2021 — 23:36

TÓPICOS
- COVID-19
- plano nacional de vacinação
- EUA
- Johnson & Johnson

Relacionados

COVID-19
**Agência Europeia está em contacto com autoridades americanas**

COVID-19
**Johnson & Johnson vai adiar entrega das vacinas na Europa**

COVID-19
**EUA recomendam pausa na utilização da vacina da Johnson & Johnson após casos de coágulos**

O anúncio foi feito esta terça-feira e teve o efeito de um balde de água fria. Mais um contratempo para o processo de vacinação na União Europeia (UE). A entrega da vacina da Jonhson & Jonhson, a única que chegará ao mercado em dose única e da qual Portugal esperava receber até esta quarta-feira mais de 30 mil doses, vai ser atrasada para a Europa.

**Por agora, não se sabe quanto tempo mais, apenas que este tempo está associado a mais estudos e investigação sobre a formação de coágulos notificados em seis mulheres jovens que receberam a vacina nos Estados Unidos da América. A vacina está suspensa, mas já foram administradas 6,8 milhões de doses.**

Depois deste anúncio, o DN contactou a task-force para saber de que forma é que a situação iria afetar o Plano Nacional de Vacinação contra a covid-19 e se estariam a ser preparadas alternativas para atenuar o seu impacto. E a resposta veio no sentido de que as metas definidas anteriormente se mantêm, ou seja, este atraso não deverá mexer com o plano definido, que agora assenta no critério da idade por ordem decrescente.

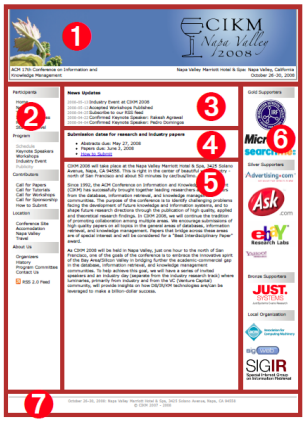


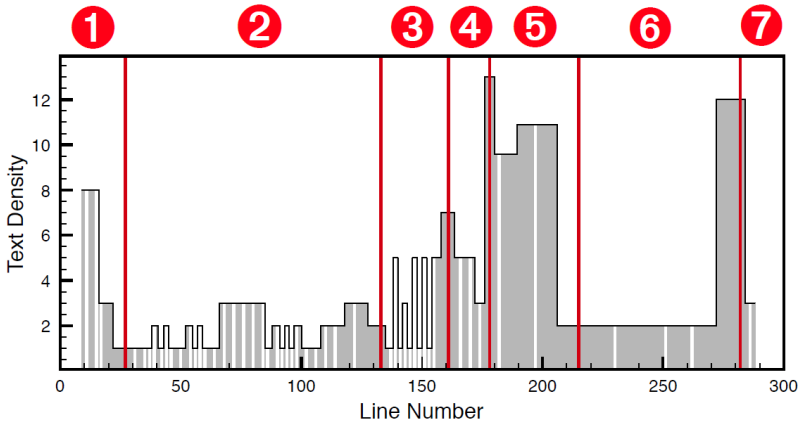
**"O Plano de Vacinação é constantemente adaptado ao fornecimento de vacinas, mas mantemos a meta inicial de ter 70% da população vacinada até ao final de verão." Ou seja, o atraso na entrega da quarta vacina a chegar ao**

```
<!-- [1.2.0.29] -->

<!DOCTYPE html>
<html lang="pt" xmlns:fb="https://www.facebook.com/2008/fbml">
<head prefix="og: https://ogp.me/ns# article: https://ogp.me/ns/article#">
    <script type="text/javascript">var _sf_startpt = (new Date()).getTime()</script>
    <meta charset="utf-8" />


<title>Atraso com vacina da Johnson n&#227;o altera metas de plano portugu&#234;s</title>
<meta charset="utf-8" />
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />
<meta property="fb:app_id" content="352742014914127" />

<meta property="ngx:content-id" content="13567270" />
<meta name="HandheldFriendly" content="true" />
<meta name="MobileOptimized" content="320" />

<meta name="apple-mobile-web-app-capable" content="yes" />
<meta name="apple-mobile-web-app-status-bar-style" content="black-translucent" />

<script type="application/ld+json">{
  "@context": "https://schema.org",
  "@type": "Organization",
  "name": "Diário de Notícias",
  "url": "https://www.dn.pt/",
  "logo": {
    "@type": "ImageObject",
    "url": "https://static.globalnoticias.pt/dn/common/images/favicons/mstile-310x310.png",
    "height": 310,
```

# Web pages

- Web pages are divided into different parts (title, abstract, body, etc)
- Each part has a specific relevance to the main content
- A Web page can be divided by its HTML structure (e.g.,  tags) or by its visual aspect.



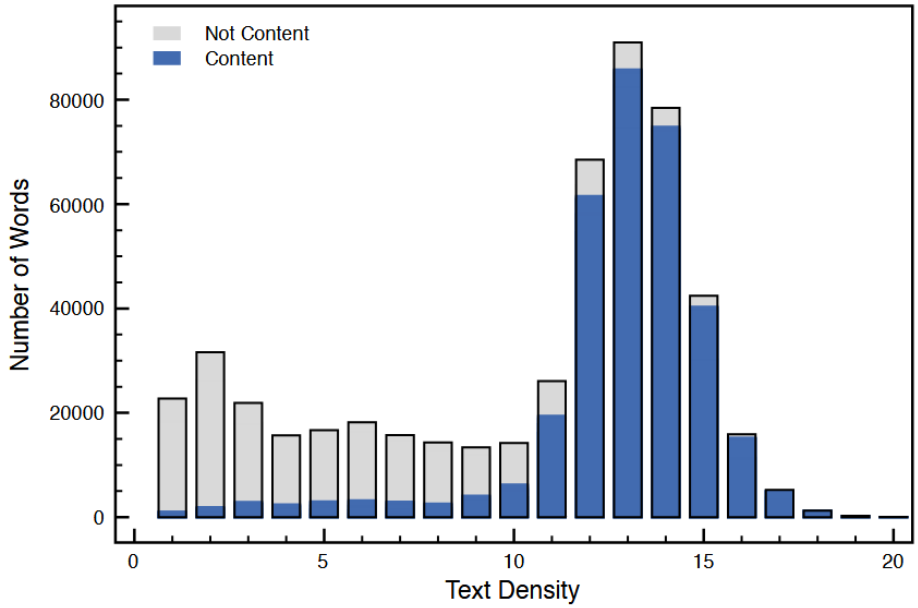



# Web page segmentation methods

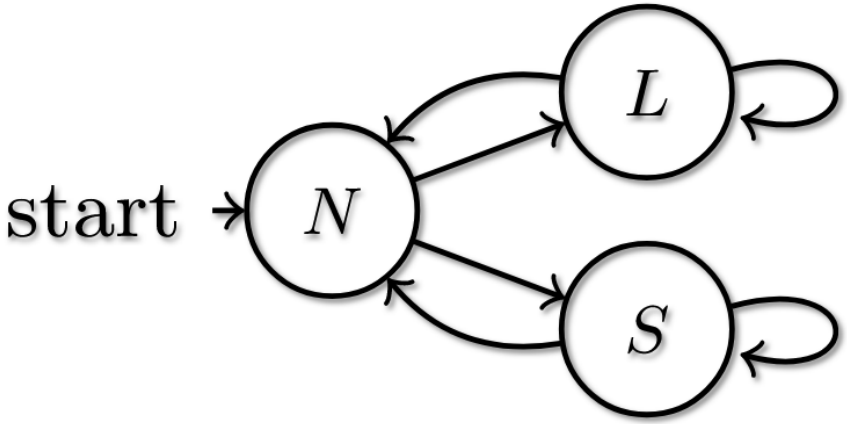
- Segmenting visually
  - Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2003). VIPS: A vision-based page segmentation algorithm.
- Densitometric approach
  - Kohlschütter, C., and Nejdl, W., (2008). A densitometric approach to web page segmentation. ACM Conference on Information and Knowledge Management (CIKM '08).
- Linguistic approach
  - Kohlschütter, C. , Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. ACM Web Search and Data Mining.

    https://boilerpipe-web.appspot.com/
    https://github.com/kohlschutter/boilerpipe
    https://github.com/jmriebold/BoilerPy3

# HTML Text extraction (Density method)



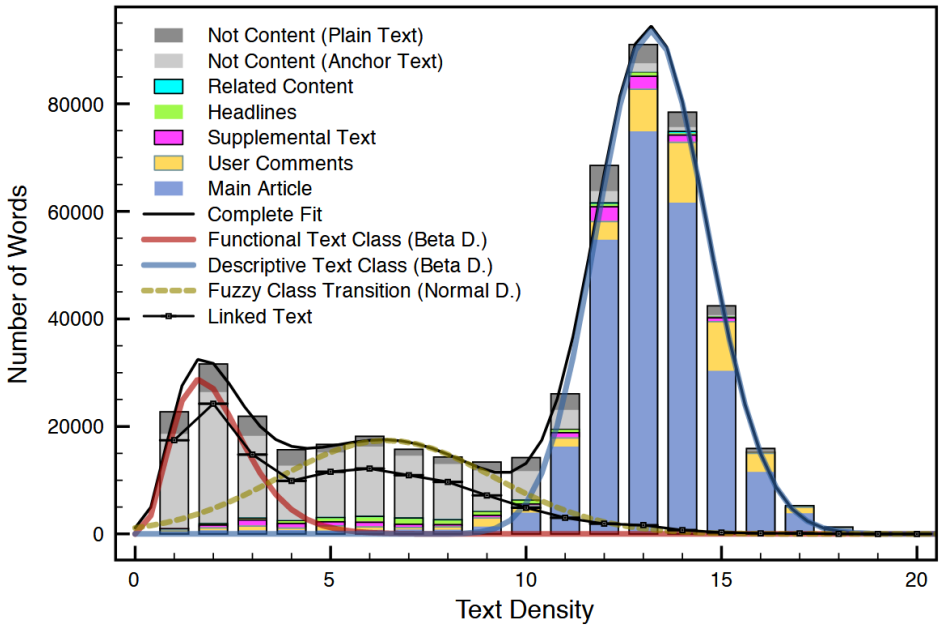
# Random writer model




L = "Long Text"
S = "Short Text"

$$P_S(N) \gg P_L(N)$$
$$P_N(L) = 1 - P_N(S)$$

# HTML Text extraction (Density + Linguistic)




https://boilerpipe-web.appspot.com/
https://github.com/kohlschutter/boilerpipe
https://github.com/jmriebold/BoilerPy3

Kohlschütter, C. , Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. ACM Web Search and Data Mining.

14 ABR 2021
Nº 55511


Índice



# Atraso com vacina da Johnson não altera metas de plano português

Task-force para Plano de Vacinação contra a Covid-19 diz que meta de ter 70% da população vacinada até ao fim do verão está de pé, apesar de a vacina da Johnson & Johnson estar agora suspensa nos EUA e ir chegar mais tarde à Europa.


Desde o início do processo de vacinação, já foram administradas mais de 6,8 milhões de doses da vacina da Johnson & Johnson nos EUA.

Ana Mafalda Inácio
13 Abril 2021 — 23:36

TÓPICOS
- COVID-19
- plano nacional de vacinação
- EUA
- Johnson & Johnson

Relacionados

COVID-19
**Agência Europeia está em contacto com autoridades americanas**

COVID-19
**Johnson & Johnson vai adiar entrega das vacinas na Europa**

COVID-19
**EUA recomendam pausa na utilização da vacina da Johnson & Johnson após casos de coágulos**

O anúncio foi feito esta terça-feira e teve o efeito de um balde de água fria. Mais um contratempo para o processo de vacinação na União Europeia (UE). A entrega da vacina da Jonhson & Jonhson, a única que chegará ao mercado em dose única e da qual Portugal esperava receber até esta quarta-feira mais de 30 mil doses, vai ser atrasada para a Europa.

**Por agora, não se sabe quanto tempo mais, apenas que este tempo está associado a mais estudos e investigação sobre a formação de coágulos notificados em seis mulheres jovens que receberam a vacina nos Estados Unidos da América. A vacina está suspensa, mas já foram administradas 6,8 milhões de doses.**

Depois deste anúncio, o DN contactou a task-force para saber de que forma é que a situação iria afetar o Plano Nacional de Vacinação contra a covid-19 e se estariam a ser preparadas alternativas para atenuar o seu impacto. E a resposta veio no sentido de que as metas definidas anteriormente se mantêm, ou seja, este atraso não deverá mexer com o plano definido, que agora assenta no critério da idade por ordem decrescente.



**"O Plano de Vacinação é constantemente adaptado ao fornecimento de vacinas, mas mantemos a meta inicial de ter 70% da população vacinada até ao final de verão." Ou seja, o atraso na entrega da quarta vacina a chegar ao**

# Boilerpipe output example

http://boilerpipe-web.appspot.com/
https://www.dn.pt/edicao-do-dia/14-abr-2021/atraso-com-vacina-da-johnson-nao-altera-metas-de-plano-portugues-13567270.html

Atraso com vacina da Johnson não altera metas de plano português

Task-force para Plano de Vacinação contra a Covid-19 diz que meta de ter 70% da população vacinada até ao fim do verão está de pé, apesar de a vacina da Johnson & Johnson estar agora suspensa nos EUA e ir chegar mais tarde à Europa.

covid-19 EUA recomendam pausa na utilização da vacina da Johnson & Johnson após casos de coágulos

Desde o início do processo de vacinação, já foram administradas mais de 6,8 milhões de doses da vacina da Johnson & Johnson nos EUA.

O anúncio foi feito esta terça-feira e teve o efeito de um balde de água fria. Mais um contratempo para o processo de vacinação na União Europeia (UE). A entrega da vacina da Jonhson & Jonhson, a única que chegará ao mercado em dose única e da qual Portugal esperava receber até esta quarta-feira mais de 30 mil doses, vai ser atrasada para a Europa.

Por agora, não se sabe quanto tempo mais, apenas que este tempo está associado a mais estudos e investigação sobre a formação de coágulos notificados em seis mulheres jovens que receberam a vacina nos Estados Unidos da América. A vacina está suspensa, mas já foram administradas 6,8 milhões de doses.

Depois deste anúncio, o DN contactou a task-force para saber de que forma é que a situação iria afetar o Plano Nacional de Vacinação contra a covid-19 e se estariam a ser preparadas alternativas para atenuar o seu impacto. E a resposta veio no sentido de que as metas definidas anteriormente se mantêm, ou seja, este atraso não deverá mexer com o plano definido, que agora assenta no critério da idade por ordem decrescente .

"O Plano de Vacinação é constantemente adaptado ao fornecimento de vacinas, mas mantemos a meta inicial de ter 70% da população vacinada até ao final de verão." Ou seja, o atraso na entrega da quarta vacina a chegar ao mercado europeu e nacional, pelos vistos, não irá interferir nas metas a atingir. Isto apesar do facto de 1,25 milhões das 2,8 milhões de doses previstas para aplicação na segunda fase de vacinação, a qual já começou em muitos concelhos, deverem ser da Jonhson & Jonhson e chegar até ao final de julho. Até ao final do ano de 2021, são esperadas um total de 4,5 milhões de doses desta vacina.

A ministra da Saúde, questionada sobre o assunto à saída da reunião com os peritos, no Infarmed, considerou ser cedo para comentar a recomendação emitida ontem pelas autoridades de saúde norte-americanas, no sentido de se fazer uma pausa na administração da vacina da Janssen. "Recebemos essa notícia enquanto decorria a reunião. Neste momento é prematuro falar. Essa vacina não foi ainda iniciada na UE, estamos a receber as primeiras doses esta semana. Os efeitos adversos estão descritos, mas no balanço risco-benefício continuamos a saber que a vacina é a nossa melhor arma para sairmos desta doença", afirmou.

Entrega da vacina depende de estudos a coágulos

O atraso na entrega à Europa está dependente dos resultados dos estudos e da avaliação que serão feitos pelo Centro de Controlo de Doenças dos EUA e da FDA (Food Drugs e Administration), organismo regulador do medicamento daquele país, sobre os casos de formação de coágulos detetados em pessoas que foram vacinadas com o produto da Johnson & Johnson.

Estes dois organismos anunciaram ontem estarem a investigar este tipo de efeito adverso notificado em seis mulheres nos dias a seguir a terem sido vacinadas, em combinação com contagens de plaquetas reduzidas. Um anúncio conjunto que levou o próprio laboratório, Jassen, a decidir que só entregaria vacinas em outros países após os resultados dos estudos que vão ser feitos por estes organismos.

Entretanto, a Agência Europeia do Medicamento (EMA), que aprovou esta vacina a 11 de março, também já anunciou estar em contacto com as autoridades americanas e a "investigar" os eventos tromboembolismo notificados após a administração desta vacina. "No contexto desta revisão, o comité de segurança da EMA (PRAC) está a investigar todos os casos relatados e decidirá se será necessária ou não uma ação regulatória", apontaram os peritos do Comité de Segurança da EMA, sublinhando que atualmente não está claro haver uma associação causal entre a vacinação e a manifestação de tais fenómenos.

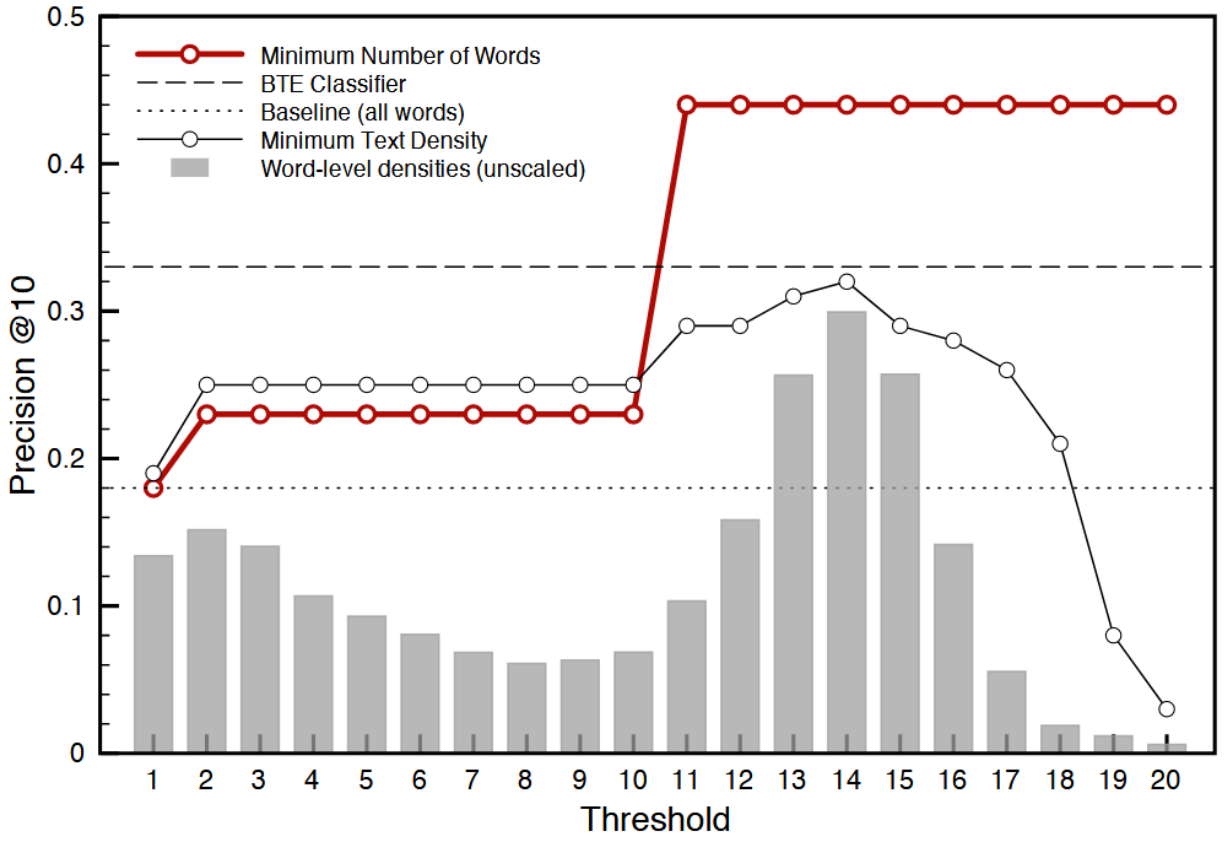
Contactado pelo DN, o porta-voz da Agência afirma que os peritos "estão a par de que o uso da vacina da Janssen para Covid-19 foi interrompido nos EUA", durante a realização de novas avaliações pelas autoridades americanas dos "dados de seis casos relatados nos EUA de um tipo raro e grave de coágulo sanguíneo em indivíduos que receberam a vacina".

O porta-voz referiu ainda que "a EMA está em contacto com o FDA dos EUA e outros reguladores internacionais sobre o assunto [e] comunicará posteriormente após a conclusão da avaliação" , esperando que "os pareceres científicos da Agência forneçam aos Estados-Membros as informações de que necessitam para tomarem decisões sobre a utilização de vacinas nas suas campanhas nacionais de vacinação".

Há uma semana a Europa estava a braços com o mesmo problema relativamente à vacina da AstraZeneca. A EMA suspendeu a sua administração para avaliar os casos de formação de coágulos detetados em pessoas de vários países, nomeadamente no Reino Unido, nas duas semanas seguintes à toma desta vacina. O comité de segurança considerou haver uma possível ligação entre a tomada da vacina e este efeito adverso e os países decidiram rever a sua administração. Portugal impõe a sua aplicação a maiores de 60 anos.

Com João Francisco Guerreiro, em Bruxelas

# Performance



# HTML information extraction

- BoilerPlate:
  - http://www.l3s.de/~kohlschuetter/boilerplate/
- BeatifulSoup:
  - https://www.crummy.com/software/BeautifulSoup/

# Web Document Categorization

**HTML documents**

**Uncategorized text documents**

**Taxonomy**


Parser

BoW

Classifier

Economy

Recipes

Sports

...

# Text pre-processing

- Instead of aiming at fully understanding a text document, IR takes a pragmatic approach and looks at the most elementary textual patterns
  - e.g. a simple histogram of words, also known as “bag-of-words”.
- Heuristics capture specific text patterns to improve search effectiveness
  - Enhances the simplicity of word histograms
- The most simple heuristics are stop-words removal and stemming

# Character processing and stop-words

- Term delimitation
- Punctuation removal
- Numbers/dates
- Stop-words: remove words that are present in all documents
  - *a, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, such, that, the, their, then, there, these, they, this, to, was, will…*

# Stemming and lemmatization

- Stemming: Reduce terms to their "roots" before indexing
  - "Stemming" suggest crude affix chopping
  - **e.g., automate(s), automatic, automation all reduced to automat.**
    - http://tartarus.org/~martin/PorterStemmer/
    - http://snowball.tartarus.org/demo.php

- Lemmatization: Reduce inflectional/variant forms to base form, e.g.,
  - *am, are, is* $\rightarrow$ *be*
  - *car, cars, car's, cars'* $\rightarrow$ *car*

# N-grams

- An n-gram is a sequence of items, e.g. characters, syllables or words.
- Can be applied to text spelling correction
  - *“interactive meida” >>>> “interactive media”*
- Can also be used as indexing tokens to improve Web page search
    - You can order the Google n-grams (6DVDs):
        - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
- N-grams were under some criticism in NLP because they can add noise to information extraction tasks
  - ...but are widely successful in IR to infer document topics.

# Documents representation

- Parsing all documents allows the creation of a vocabulary of all the words that exist in the domain:

$$V = \{(x_1, 1), \ldots, (x_L, L)\}$$

- After the text analysis steps, a document (e.g. Web page) is represented as a vector of terms, n-grams, etc. :

$$d_i = \left(x_{i,1}, \ldots, x_{i,L}\right) \qquad x_{i,j} \geq 0$$

Number of times word $x_j$ occurs in document $d_i$

# A (simple) text parser

```
data = [("me gusta comer en la cafeteria".split(), "SPANISH"),
        ("Give it to me".split(), "ENGLISH"),
        ("No creo que sea una buena idea".split(), "SPANISH"),
        ("No it is not a good idea to get lost at sea".split(), "ENGLISH")]

test_data = [("Yo creo que si".split(), "SPANISH"),
             ("it is lost on me".split(), "ENGLISH")]
```

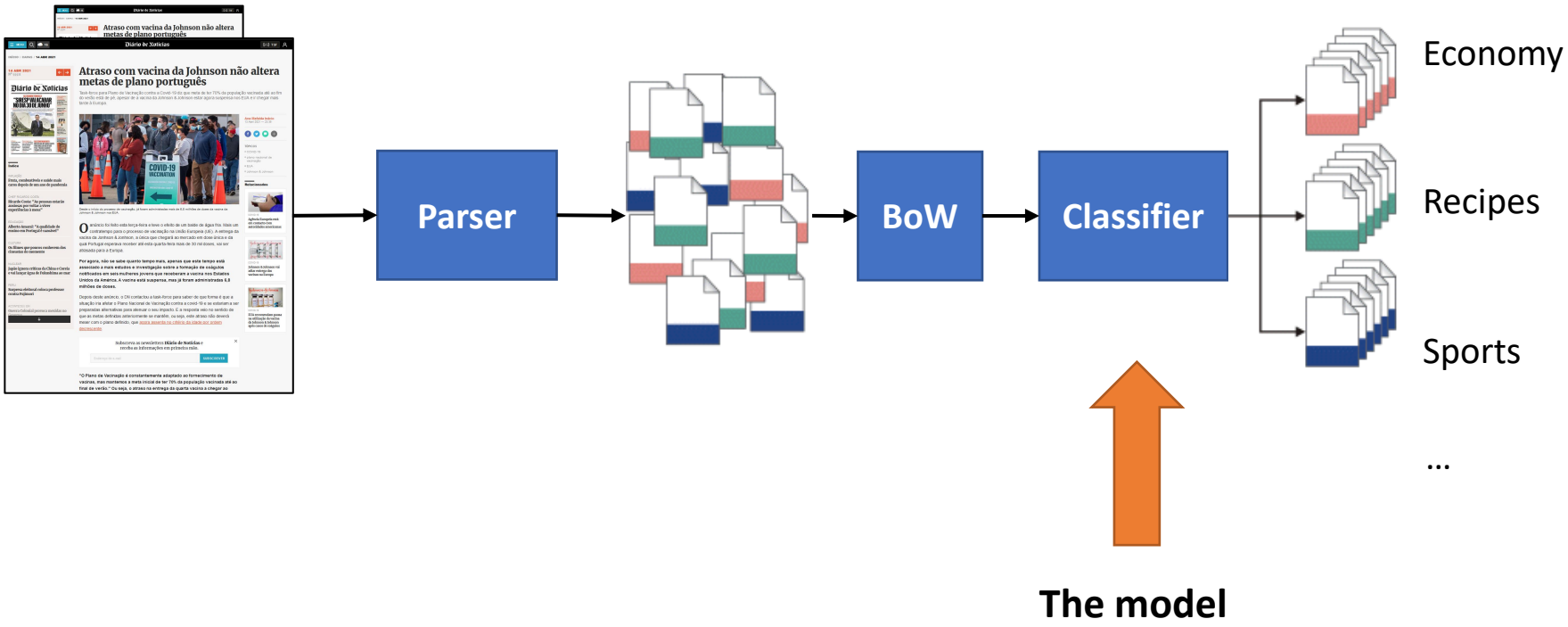
https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html

# Bag of Words representation

```
# Build a dictionary word_to_ix
# to map each word in the vocab
# to a unique integer
word_to_ix = {}
for sent, _ in data + test_data:
    for word in sent:
        if word not in word_to_ix:
            word_to_ix[word] = len(word_to_ix)
print(word_to_ix)


VOCAB_SIZE = len(word_to_ix)

# Convert a document into a BOW vector
def make_bow_vector(sentence, word_to_ix):
    vec = torch.zeros(len(word_to_ix))
    for word in sentence:
        vec[word_to_ix[word]] += 1
    return vec.view(1, -1)
```

https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html

# Web Document Categorization

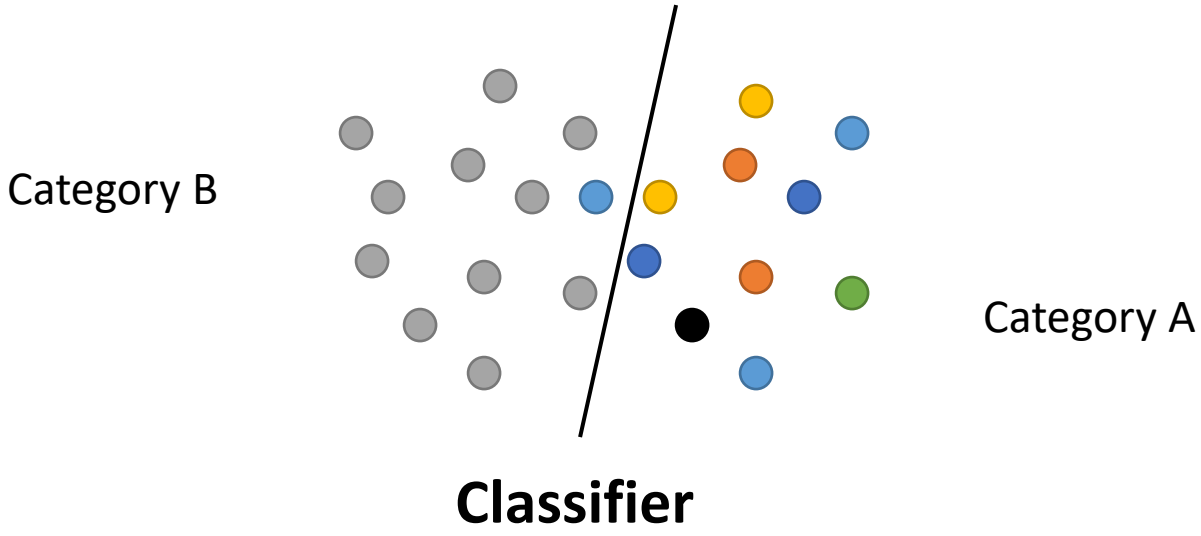


# Categorization/Classification

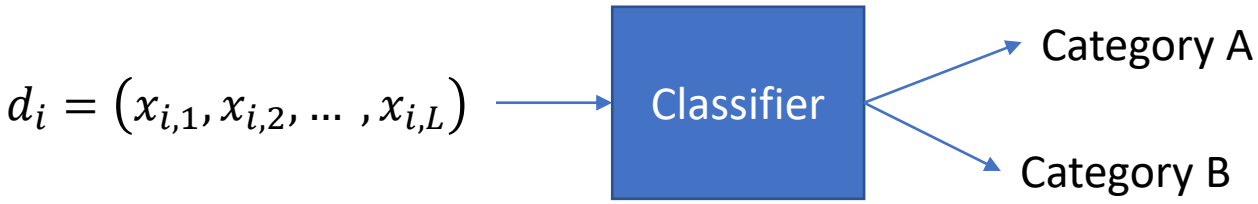
- Given:
  - A description of an instance, $d \in X$
    - $X$ is the *instance language* with vocabulary V.
      - Issue: how to represent text documents.
      - Usually some type of high-dimensional space
  - A fixed set of classes: $C = \{c_1, c_2, \ldots, c_L\}$
- Determine:
  - The category of $d$: $\gamma(d) \in C$, where $\gamma(d)$ is a *classification function* whose domain is $X$ and whose range is $C$.
    - We want to know how to build classification functions ("classifiers").

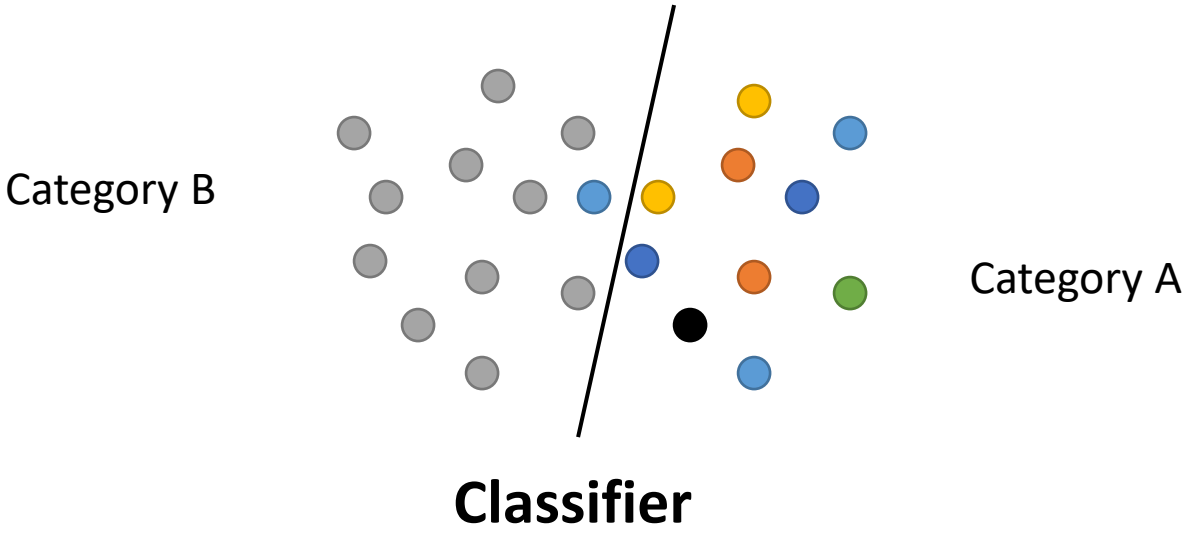
# Classification task

- For new unseen documents, we wish to classify documents with one of the known classes.
- New documents are represented in some feature space and then a machine learning algorithm classifies the new documents.



# Input sample *can be* the document word counts



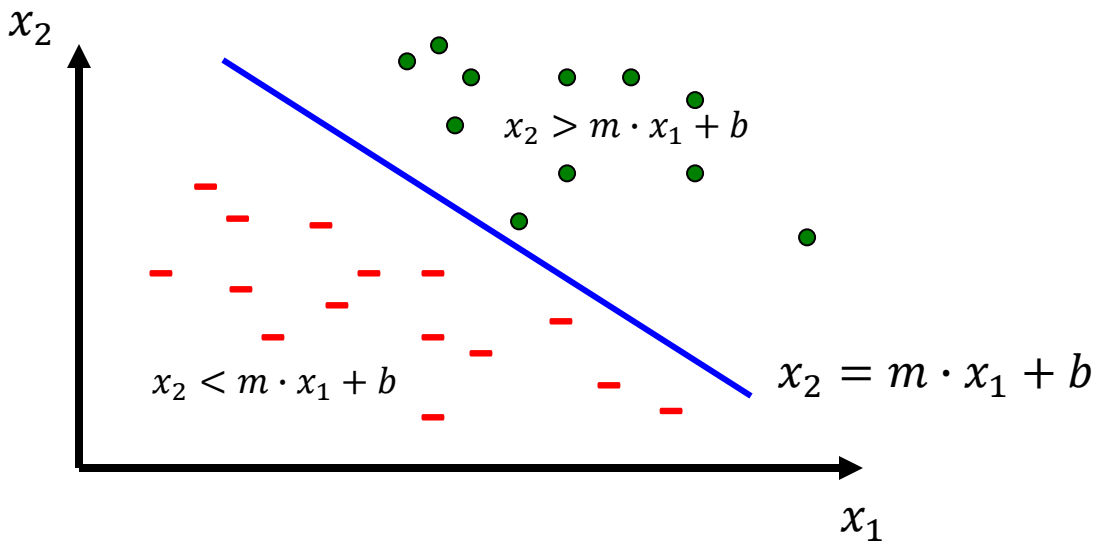



# Perceptron

- All sample vectors $\boldsymbol{x^{(j)}}$ have their corresponding label $\boldsymbol{y^{(j)} = \{+1, -1\}}$

- **The perceptron performs a binary prediction $\hat{y}$ based on the observed data $x$ :**

$$\hat{y} = f(x) = \begin{cases} +1 & , if\ x_2 \geq\ m \cdot x_1 + b \\ -1 & , if\ x_2 < m \cdot x_1 + b \end{cases}$$

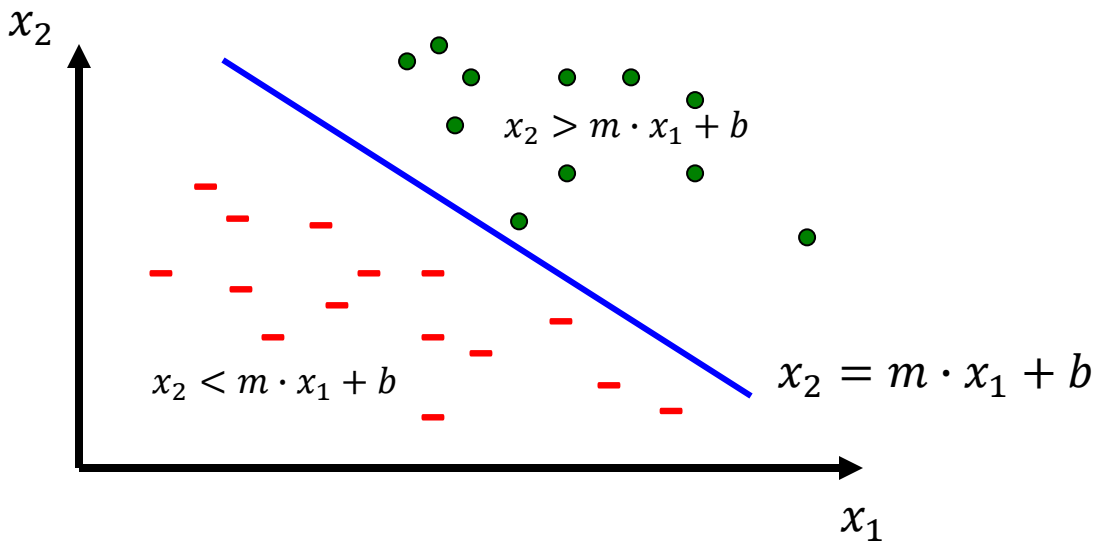


# Perceptron

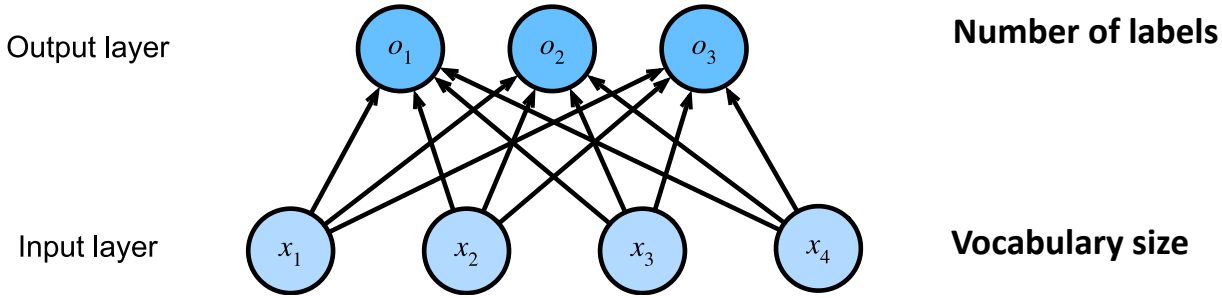
- All sample vectors $\boldsymbol{x^{(j)}}$ have their corresponding label $\boldsymbol{y^{(j)} = \{+1, -1\}}$

- **The perceptron performs a binary prediction $\hat{y}$ based on word counts of document $d_j = (x_1, \ldots, x_L)$**

$$\hat{y} = f(d_j) = \begin{cases} +1 & ,if\ 0 \geq\ m \cdot x_1 + b - x_2 \\ -1 & ,if\ 0 < m \cdot x_1 + b - x_2 \end{cases}$$



# The linear model




**Output layer** **Input layer** **Weights**

$$\begin{bmatrix} o_1 \\ o_2 \\ o_3 \end{bmatrix} = [x_1 \quad x_2 \quad x_3 \quad x_4] \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \end{bmatrix}$$

**Label of document** **Vector with words weights** **The linear model**

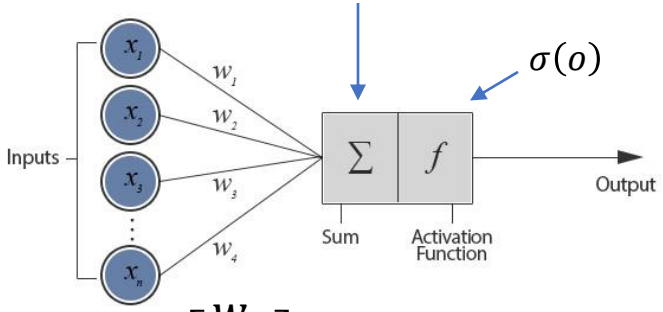
# The sigmoid activation function

- Consider only **one output**:

$$o = w_0 + w_1 x_1 + \ \ldots + w_n x_n$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \\ x_n \end{bmatrix}$$

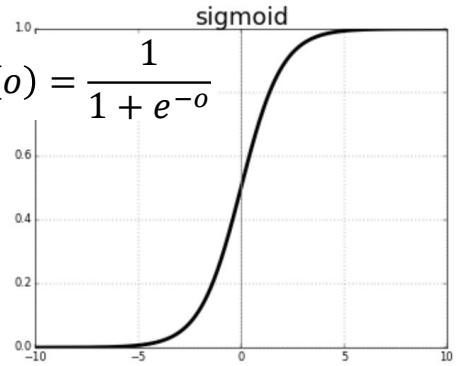
Word frequency counts




$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \\ w_n \end{bmatrix}$$

Word weights

$$y = \sigma(o) = \frac{1}{1 + e^{-o}}$$




$$o = w_0 + w_1 x_1 + \ \ldots + w_n x_n$$

# The softmax activation function

- The softmax function is the generalization of the sigmoid function to multiple classes

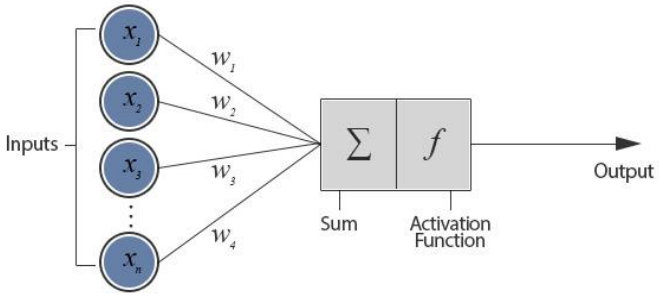
$$\text{softmax}(o_k) = y_k = \frac{e^{o_k}}{e^{o_1} + \cdots + e^{o_k} + \cdots + e^{o_L}} = \frac{e^{o_k}}{\sum_i e^{o_i}}$$

- It's appropriate when we wish to compute the joint probability of all classes:

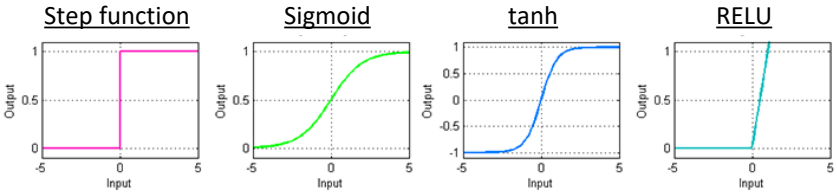
$$\text{softmax}([o_1, o_2, o_3]) = [y_1, y_2, y_3] = \left[ \frac{e^{o_1}}{\sum_i e^{o_i}} \quad \frac{e^{o_2}}{\sum_i e^{o_i}} \quad \frac{e^{o_3}}{\sum_i e^{o_i}} \right]$$

# Activation functions

- The perceptron was initially proposed with the step function.
- Historically, other activation functions have been studied.
- The perceptron with the sigmoid activation function corresponds to the logistic regression model.

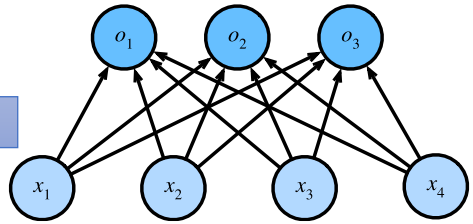



**Activation functions**

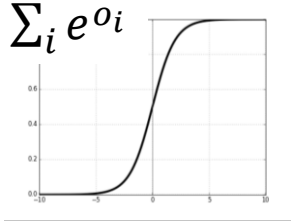


# The linear classifier model

```
class BoWClassifier(nn.Module):

    def __init__(self, num_labels, vocab_size):
        # calls the init function of nn.Module.
        super(BoWClassifier, self).__init__()

        # Define the parameters that you will need
        self.linear = nn.Linear(vocab_size, num_labels)


    def forward(self, bow_vec):
        # Pass the input through the linear layer,
        # then pass that through log_softmax.
        return softmax(self.linear(bow_vec), dim=1)
```

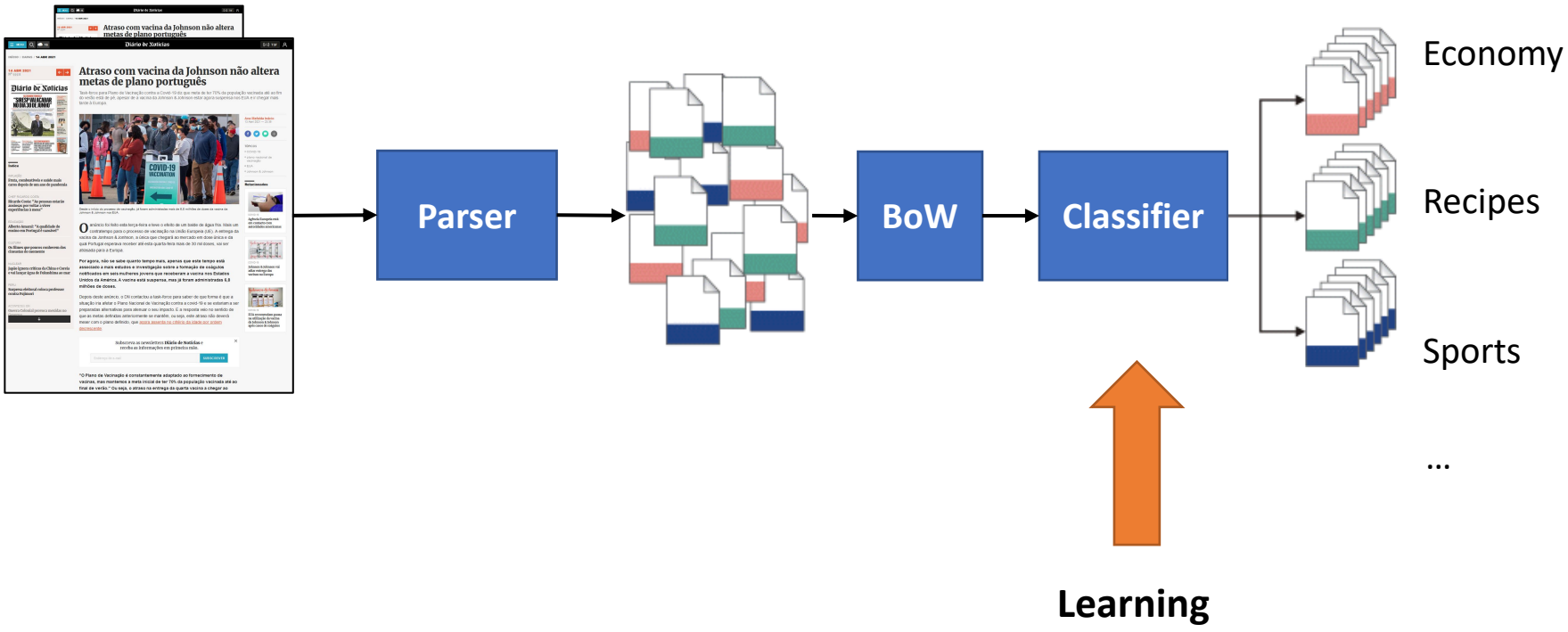



$$y_k = \frac{e^{o_k}}{\sum_i e^{o_i}}$$



https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html

# Web Document Categorization

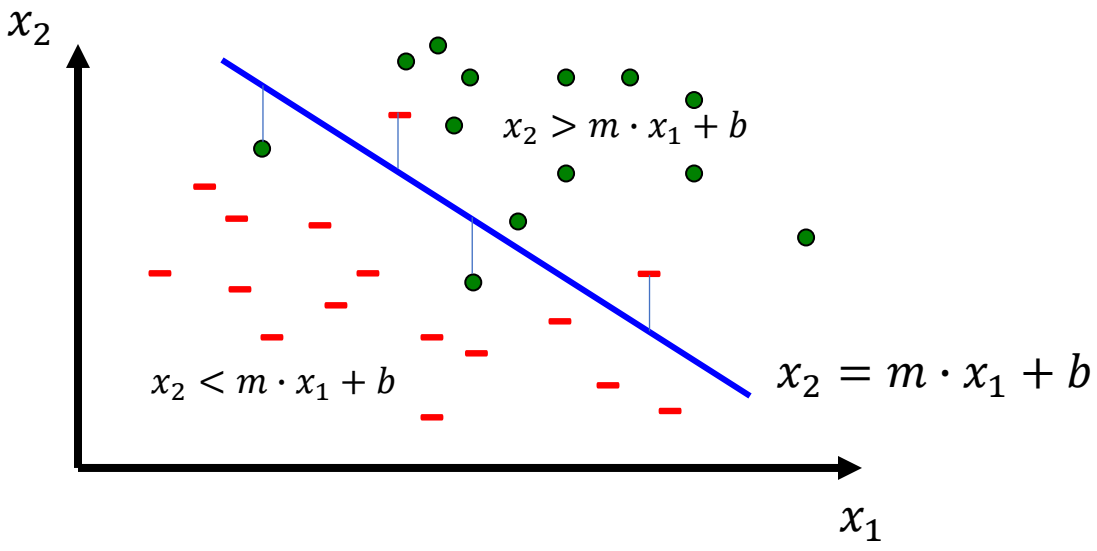


# Model error

- The Mean Square Error (MSE) measures the error between the true labels and the predicted labels

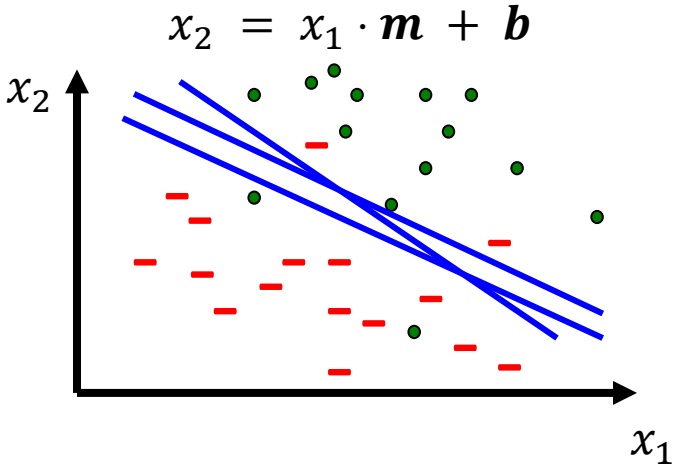
$$MSE = \frac{1}{N}\sum_i^N (error_i)^2$$

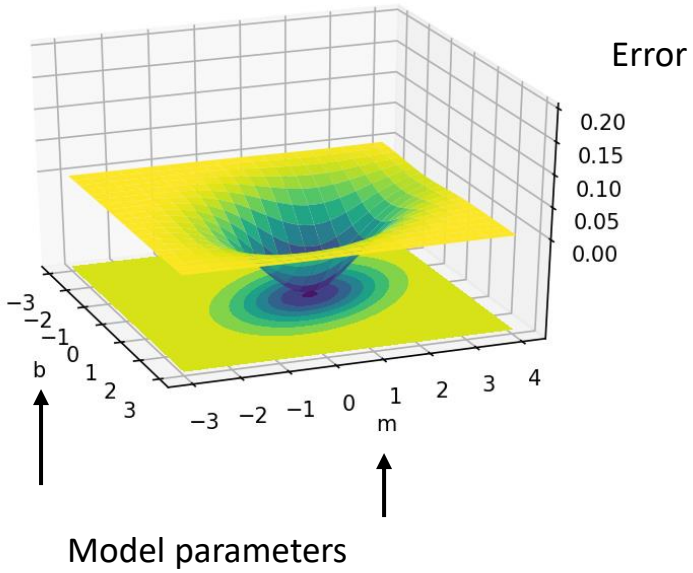
$$error_i = true\_label_i - predictedLabel_i$$



# Minimizing the error

$$MeanSquareError = \frac{1}{N}\sum_{i}^{N}(label_i - predictedLabel_i)^2$$



# Loss functions and optimizer

- There are multiple error functions
  - Mean squared error loss
  - Binary cross entropy
  - negative log likelihood loss

- There are multiple optimization methods:

```
loss_function = nn.NLLLoss()
optimizer = optim.SGD(model.parameters(), lr=0.1)
```

https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html

# Training loop – Stochastic Gradient Descent

```
loss_function = nn.NLLLoss()
optimizer = optim.SGD(model.parameters(), lr=0.1)

for epoch in range(100):
    for instance, label in data:
        # Step 1. Get the document vector
        bow_vec = make_bow_vector(instance, word_to_ix)

        # Step 2. Get the document label
        target = make_target(label, label_to_ix)

        # Step 3. Run our forward pass
        log_probs = model(bow_vec)

        # Step 4. Compute the loss for all documents
        loss = loss_function(log_probs, target)

        # Step 5. Compute the gradients for all parameters
        loss.backward()

        # Step 6. Update the parameters
        optimizer.step()

        # Step 7. Reset the gradients
        model.zero_grad()
```

Doc text -> $(bow)$

Doc label -> $(true_y)$

$pred_y = model(bow)$

$loss(true_y, pred_y)$

$\frac{\partial loss(true_y, pred_y)}{\partial W}$

$W_{n+1} = W_n + \gamma \cdot \frac{\partial loss(true_y, pred_y)}{\partial W}$

https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html

# Multiple labels

- Each document can have more than one label simultaneously
  - Multiple sigmoids
- Each document has a distribution of labels
  - Softmax
- Each document has only one label and it excludes the others
  - Softmax

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

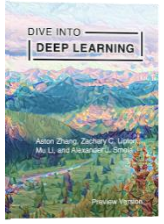
$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

# Summary

- Web text extraction
- Text processing and Bag of Words
- The Perceptron
- Model learning

- PyTorch introduction
  - https://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html

- References:




Aston Zhang, Zachary Lipton, Mu Li, and Alex Smola, “Dive into Deep Learning” Chapter 3

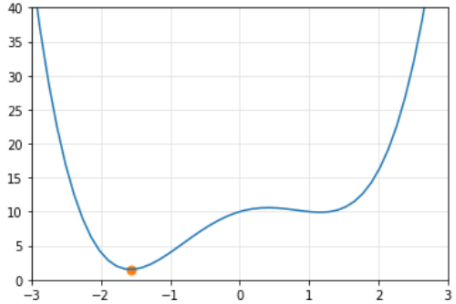
# Notes on learning, model selection and evaluation

João Magalhães

# Learning a model parameters

- Consider dataset $\mathcal{D}$ and predictive model with parameters $\beta = \{m, b\}$
- The goal is to find the parameters that minimize the difference between predictions and ground-truth:

$$loss(\beta, \mathcal{D}) = \frac{1}{N}\sum_{i}^{N}(label_i - prediction_i(\beta, d_i))^2$$



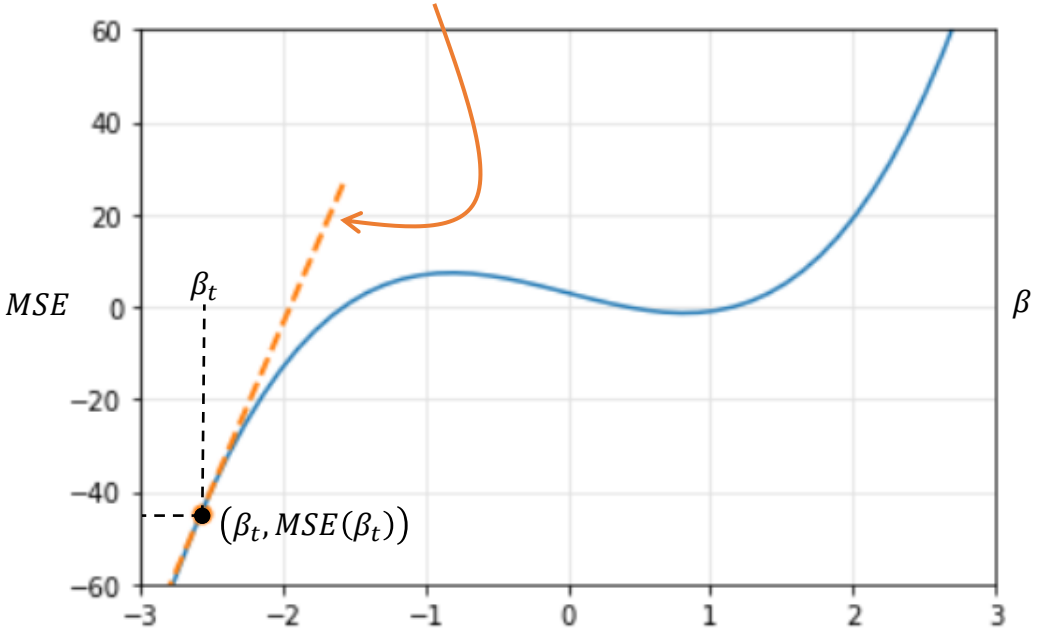

- Finding the minimum is equivalent to finding the zero of the error gradient function

$$loss'(\beta, \mathcal{D}) = \frac{\partial}{\partial\beta}\left(\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}(label_i - prediction_i(\beta, d_i))^2\right)$$

# Newton method

- The Newton method finds the zero of a function by using its tangent, i.e. the gradient function, on each point of the function :

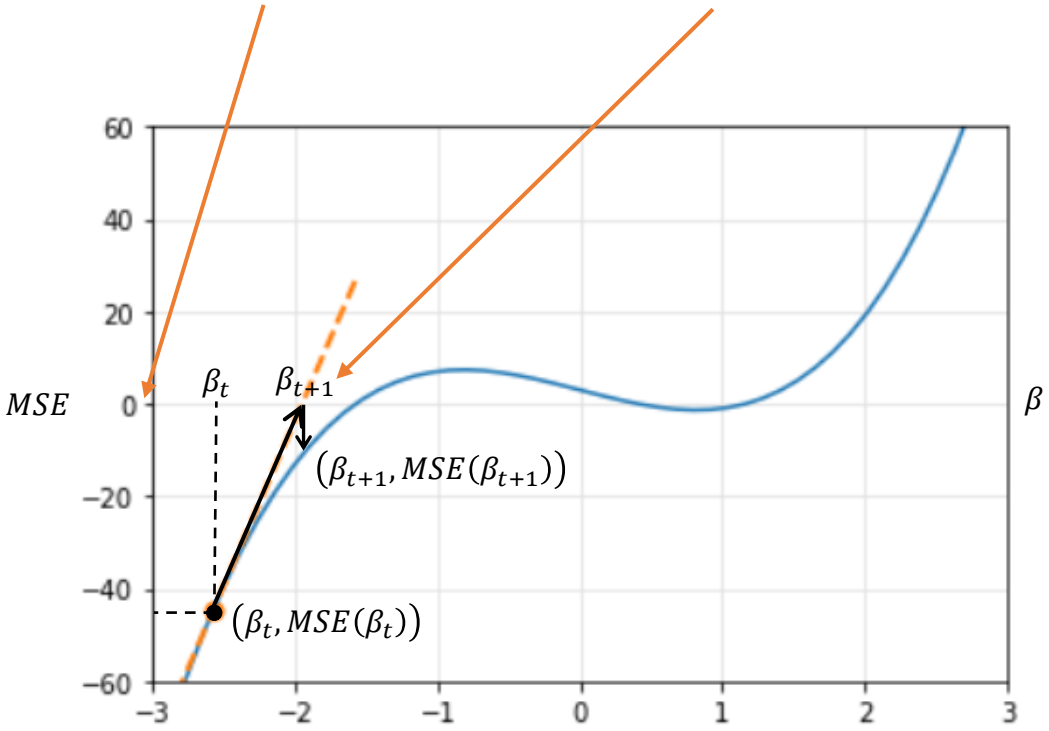
$$\underbrace{loss(\beta \ \ ,\mathcal{D})}_{y} = \underbrace{loss'(\beta_t,\mathcal{D})}_{m} \cdot \underbrace{(\beta - \beta_t)}_{x} + \underbrace{loss(\beta_t,\mathcal{D})}_{b}$$



# Newton method

- Iteratively, we find the zero of the tangent and the new parameters $\beta_{t+1}$ are taken from this point:

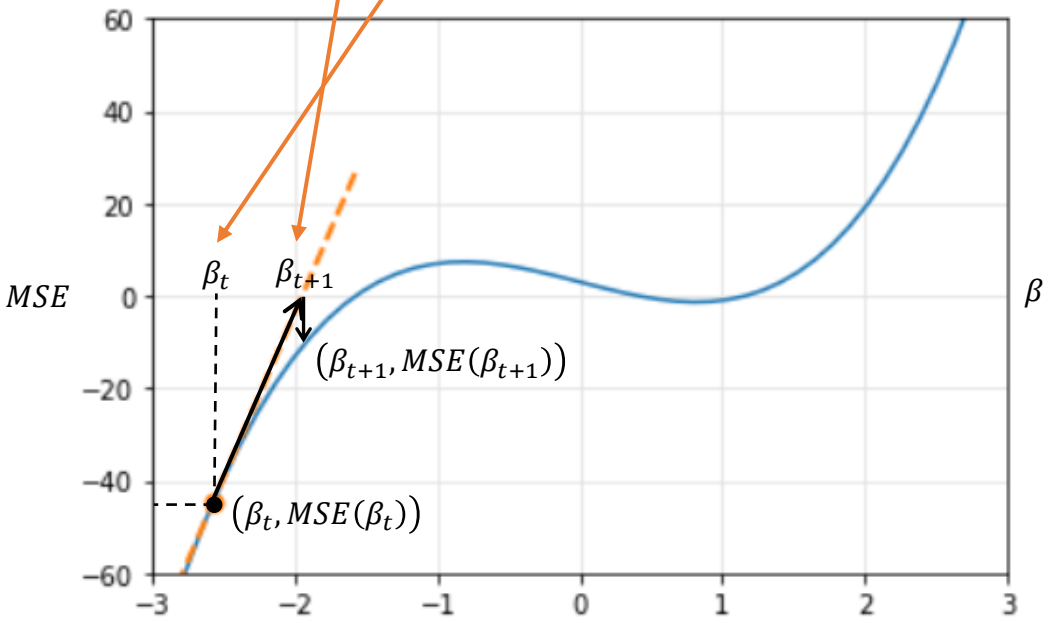
$$loss(\beta_{t+1}, \mathcal{D}) = 0 = loss'(\beta_t, \mathcal{D}) \cdot (\beta_{t+1} - \beta_t) + loss(\beta_t, \mathcal{D})$$



# Newton method

- Solving the previous expression in order to $\beta_{t+1}$ is now straightforward:

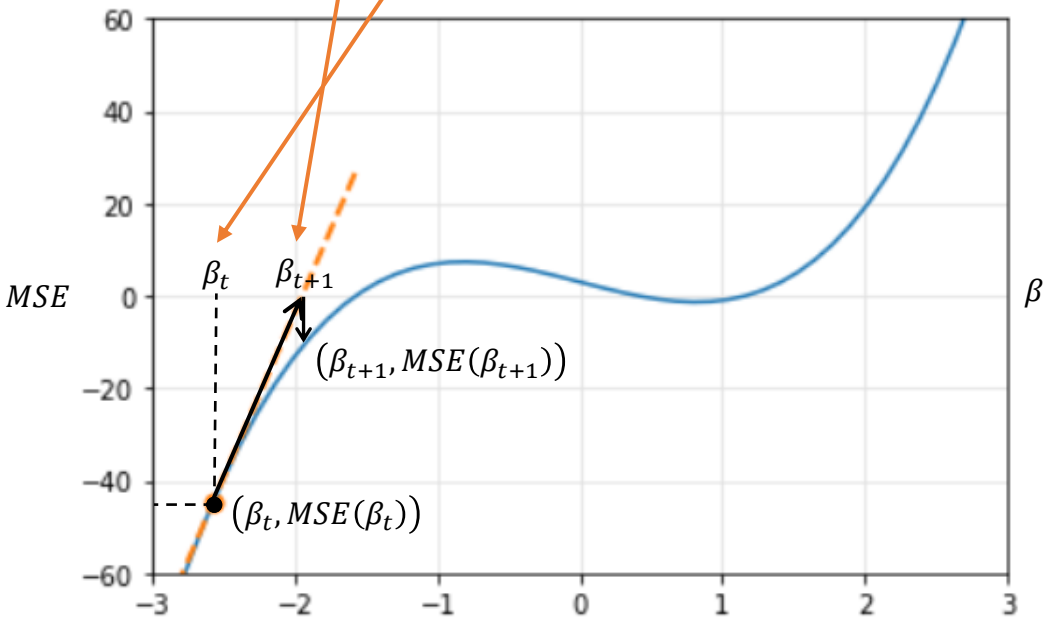
$$\beta_{t+1} = \beta_t - \frac{loss(\beta_t, \mathcal{D})}{loss'(\beta_t, \mathcal{D})}$$



# Newton method

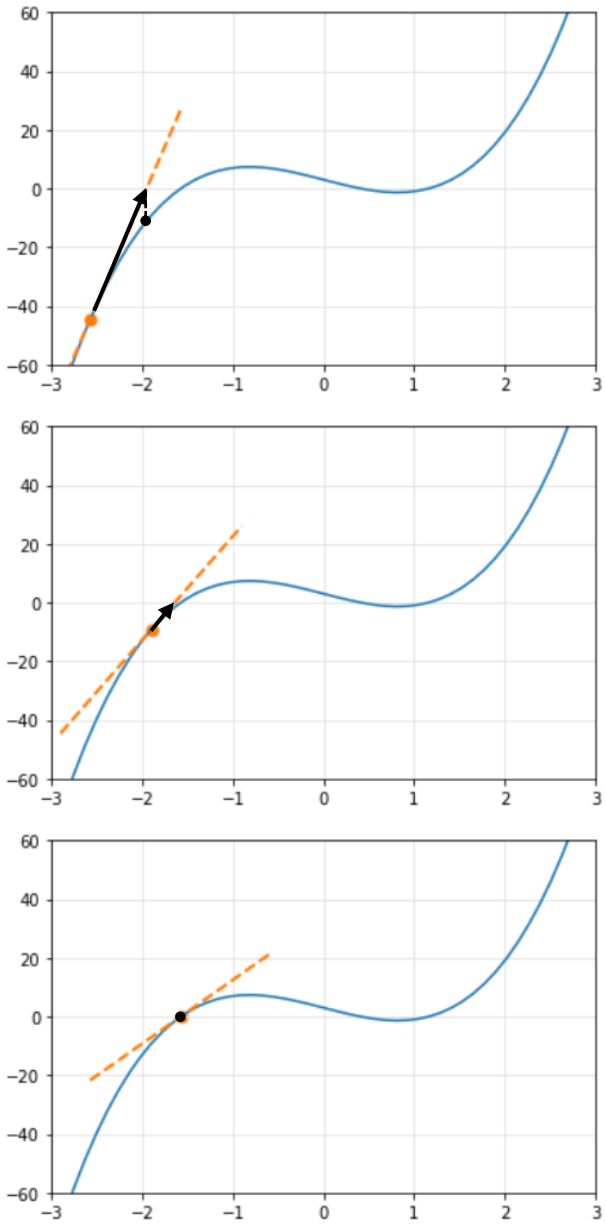
- Solving the previous expression in order to $\beta_{t+1}$ is now straightforward:

$$\beta_{t+1} = \beta_t - \frac{loss(\beta_t, \mathcal{D})}{loss'(\beta_t, \mathcal{D})}$$



# Newton method

- Iteratively, the Newton method uses the tangent to the function on its last approximation.
- The x value corresponding to the intercept of the tangent (y=0) will be used in the next iteration.
- The algorithm stops when the function value is close enough to zero, i.e., is below a given threshold.



# Newton method for minimum of functions

- In our case, the loss function $loss(\beta_t, \mathcal{D})$ can never reach zero, so in practice we wish to find the minimum of the gradient of the error function $loss'(\beta_t, \mathcal{D})$

- This leads to:

$$\beta_{t+1} = \beta_t - \frac{loss'(\beta_t, \mathcal{D})}{loss''(\beta_t, \mathcal{D})}$$

- Unfortunately, computing the second derivative of the error function is seldomly possible and too computationally can be too complex.
  - The Gradient Descent Algorithm and its most popular variant the Stochastic Gradient Descent Algorithm are usually used for this purpose.

# Gradient Descent

- Without computing the second derivative, the Gradient Descent method replaces it by an upper bound constant, the so-called learning rate:

$$\beta_{t+1} = \beta_t - lr \cdot loss'(\beta_t, \mathcal{D})$$

- The family of Gradient Descent methods use only the first order derivative to descend the error function towards the local minimum.

- In both Newton and GD methods, the value of the MSE gradient is computed over the entire dataset.
  - This leads to very accurate estimations, but to extremely slow updates.

# Stochastic Gradient Descent

- Stochastic Gradient Descent estimates a very course approximation to the loss gradient by computing this estimate based on only one sample.

$$\beta_{t+1} = \beta_t - lr \cdot loss'(\beta_t, d_i)$$

- Although the estimate is very course, it is extremely cheap and after several updates the algorithm converges asymptotically.

# Stochastic Mini-Batch Gradient Descent

- A trade-off between GD and SGD is struck when the loss gradient is estimated over a sufficiently large batch of samples:

$$\beta_{t+1} = \beta_t - lr \cdot \sum_{i \in batch} loss'(\beta_t, d_i)$$

- This estimate is more accurate and still cheap to be computed several times over the entire dataset, thus leading to a faster convergence then with pure GD.

# SGD with Momentum

- The rationale of momentum is to soften the current gradient $loss'(\beta_t, batch)$ with past values $m_{t-1}$ of the gradient:

$$m_t = \delta \cdot m_{t-1} + loss'(\beta_t, batch_t)$$

- The parameter $\delta$ is the decaying penalty over past gradients.

- Hence, the parameters are now updated with smoothed value of the gradient:

$$\beta_{t+1} = \beta_t - lr \cdot m_t$$

# SGD with Adaptive Moment Estimation

- **ADAM is currently the most effective method to train a machine learning model.**
  - It modifies the learning rate of each parameter throughout the training iterations.

- It keeps a memory of past gradients to compute the first momentum (mean of past gradients) and the second momentum (variance of past gradients)

$$\beta_{t+1} = \beta_t - lr \frac{1}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$