

Grado en Ingeniería Informática

Explotación de la Información Módulo 4. Clasificación y Agrupamiento de Información

Antonio Ferrández Rodríguez



UNIVERSIDAD DE ALICANTE



Grupo de Procesamiento
del Lenguaje y Sistemas
de Información

1

Índice

0. Introducción

1. Sistemas de clasificación de información
2. Clasificación basada en vocabulario
3. Clasificación utilizando árboles de decisión
4. Clasificación utilizando sistemas de reglas
5. Problema del *overfitting*. Sistemas de poda
6. Part of speech tagging
7. Sistemas de agrupamiento de información
8. Sistemas de agrupamiento de información en la Recuperación de Información
9. Sistemas de agrupamiento de información particionales (algoritmo *k-mean*)
10. Sistemas de agrupamiento de información jerárquicos
11. Herramientas

Explotación de la información, Clasificación y Agrupamiento de Información

2

2

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Aprendizaje automático (*machine learning*) (https://www.w3schools.com/ai/ai_machine_learning.asp):

Traditional Computing

Data + Computer Algorithm = **Result**

Machine Learning

Data + Result = **Computer Algorithm**

3

3

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Aprendizaje automático (cont.) (https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico):

- Subcampo de las ciencias de la computación y una rama de la inteligencia artificial
- Objetivo es desarrollar técnicas que permitan que las computadoras aprendan
- Taxonomía:
 - Aprendizaje supervisado
 - Aprendizaje NO supervisado
 - Aprendizaje por refuerzo
- Problemas a resolver: predicción, clasificación y agrupamiento

4

4

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Aprendizaje automático (cont.):

- Aprendizaje supervisado
 (<https://www.nltk.org/book/ch06.html>):

The diagram illustrates the supervised learning process in two parts:

(a) Training: An input (represented by a document icon) is processed by a feature extractor to produce features (represented by a vector of boxes). These features are fed into a machine learning algorithm, which also takes a label as input to learn a model.

(b) Prediction: A new input is processed by the same feature extractor to produce features. These features are then fed into a classifier model, which outputs a predicted label.

5

5

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción


**Aprende
je
automático
o (cont.)**
 (<https://www.w3schools.com/ai/sciences.asp>):

The diagram shows a series of concentric ellipses representing the relationship between different concepts in machine learning and data science:

- Weak:** The outermost, lightest green ellipse.
- Machine Learning:** A medium green ellipse nested within Weak.
- Neural Networks:** A darker green ellipse nested within Machine Learning.
- Big Data:** A dark green ellipse nested within Neural Networks.
- Deep Learning:** A very dark green ellipse nested within Big Data.
- Strong:** The innermost, darkest green ellipse.

6

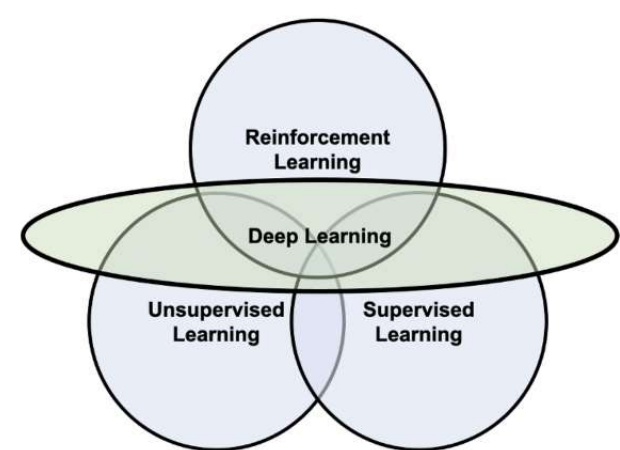
6



Explotación de la información, Clasificación y Agrupamiento de Información


0. Introducción

Aprende por refuerzo (<https://towardsai.net/article-1-introduccion-al-aprendizaje-por-refuerzo/>):



7

7

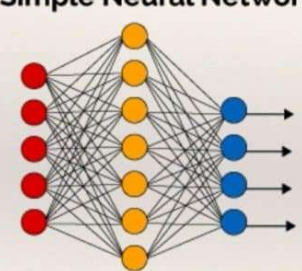


Explotación de la información, Clasificación y Agrupamiento de Información

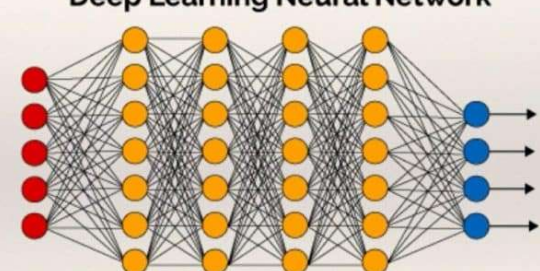
0. Introducción

Deep Learning (<https://www.iartificial.net/redes-neuronales-desde-cero-i-introduccion/>):

Simple Neural Network



Deep Learning Neural Network



● Input Layer
● Hidden Layer
● Output Layer

8

8

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Objetivo del aprendizaje automático
(<https://builtin.com/data-science/regression-machine-learning>):

Underfitting Ideal Overfitting

9

9

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

UC Irvine Machine Learning Repository:
<https://archive.ics.uci.edu/ml/index.php>

895694:		Car Evaluation
830167:		Breast Cancer Wisconsin (Diagnostic)
821347:		Wine Quality
801758:		Heart Disease
767904:		Bank Marketing
748977:		Human Activity Recognition Using Smartphones

Newest Data Sets:		
11-30-2018:		2.4 GHz Indoor Channel Measurements
11-16-2018:		Electrical Grid Stability Simulated Data
11-09-2018:		BAUM-2
11-09-2018:		BAUM-1
11-05-2018:		Parkinson's Disease Classification
11-02-2018:		Caesarian Section Classification Dataset
10-12-2018:		Superconductivity Data
10-08-2018:		Physical Unclonable Functions
10-04-2018:		Drug Review Dataset (Drugs.com)

10

10

0. Introducción

13 Deep Learning Frameworks for Natural Language Processing in Python:

<https://medium.com/@datamonsters/13-deep-learning-frameworks-for-natural-language-processing-in-python-2b84a6b6cd98>

Framework/library name	Author/Developer	License	Natural Language Processing tasks													Network types				Demo/About		
			Text classification	Text generation	Text summarization	NER (Named Entity Recognition)	POS (Part-Of-Speech) tagging	Word embeddings	Dependency parsing	SRL (Semantic Role Labeling)	Sequence tagging	Language modeling	Machine translation	Speech recognition	QA (Question Answering)/Conversations	Convolutional neural networks (CNN)	Long Short Term Memory Networks (LSTM)	Bidirectional LSTMs	Recurrent neural networks (RNN)		Recursive neural networks (RNN)	Sequence to Sequence (seq2seq) models
Chainer	Preferred Networks	MIT License	+			+		+	+						+	+	+	+	+	+		
DeepLearning4j	Adam Gibson	Apache License 2.0	+					+	+	+												https://www.youtube.com/watch?v=TjhyqAQDwZI

11


0. Introducción

Gensim (Python):

<https://pypi.org/project/gensim/>

- Gensim is a Python library for *topic modelling*, *document indexing* and *similarity retrieval* with large corpora. Target audience is the *natural language processing* (NLP) and *information retrieval* (IR) community.
- Efficient multicore implementations of popular algorithms, such as online Latent Semantic Analysis (LSA/LSI/SVD), Latent Dirichlet Allocation (LDA), Random Projections (RP), Hierarchical Dirichlet Process (HDP) or word2vec deep learning.

12




Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

- # **Deeplearning4j en Java:**
 - ▣ <https://www.youtube.com/watch?v=TjhYqAQDwZI>
- # **Entorno y lenguaje de programación R:**
 - ▣ [https://es.wikipedia.org/wiki/R_\(lenguaje_de_programaci%C3%B3n\)](https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n))
 - ▣ <https://www.r-project.org/>
 - ▣ R proporciona herramientas estadísticas (modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas
 - ▣ Repositorio oficial de paquetes desarrollados con R: <https://web.archive.org/web/20101221001753/http://cran.r-project.org/web/packages/>

13

13



Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

- # **Deep Learning en PLN:**
 - ▣ Word embeddings (word2vec, GloVe)
 - ▣ Convolutional neural networks (CNNs) para la extracción de características destacables
 - ▣ Recurrent neural networks (RNNs, LSTMs, GRUs) para modelar secuencias de texto
 - ▣ Generative adversarial networks (GANs) para generar salida en lenguaje natural humano

14

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Deep Learning en PLN (cont.):

- Charla de Chris Manning sobre *Deep Learning in NLP*:
 - https://www.youtube.com/watch?v=OQQ-W_63UgQ
 - ¿Retos del *Deep Learning* para su aplicación en PLN?
 - # Vocabularios casi infinitos
 - # Complejidad de representación, aprendizaje y uso del significado
 - # Ambigüedad: diferentes representaciones sintácticas tienen el mismo significado; humor, ironía, sentimientos, etc.
 - # Dependencias de interpretación según el mundo exterior, sentido común y conocimiento contextual
 - # Problemas lingüísticos que hay que resolver previamente: elipsis, anáfora, coordinación, yuxtaposición, ... para tener una estructura no ambigua que represente el conocimiento

15

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Deep Learning en PLN (cont.):

- Ex.Inf. Modulo 4. Cariño, he encogido a la IA que ocupe poco puede hacerla más rápida, más ecológica y más segura para la privacidad del usuario.pdf:
 - PLN tiene los modelos de IA con mayor n° de parámetros → mayor requerimiento de cálculo y potencia computacional:
 - # BERT de Google: **340 millones de parámetros**
 - # OpenAI, el generador de textos fake creíbles:
 - GPT-2: **1.500 millones de parámetros** en 40 GB
 - GPT-3: **175.000 millones de parámetros** en 700 GB ubicado en 48 GPUs de 16 GB cada una de ellas
 - # El modelo de IA creado por Nvidia: **8.300 millones de parámetros**
 - Ahora hay tendencia a reducir esos modelos de IA (Lite BERT)

16

0. Introducción

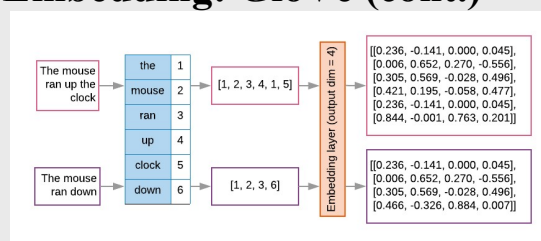
Word Embedding:

- GloVe (Stanford), Word2Vec (Google) o fastText (Facebook)
 - <https://nlp.stanford.edu/projects/glove/>
 - GloVe is an **unsupervised** learning algorithm for obtaining vector representations for words.
 - *Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space*
 - Otros métodos para generar este mapeo son **redes neuronales**, reducción de dimensionalidad en la matriz de co-ocurrencia de palabras, modelos probabilísticos, y representación explícita en términos del contexto en el cual estas palabras figuran

17

0. Introducción

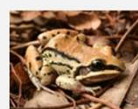
Word Embedding: GloVe (cont.)



1. Nearest neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. frog
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus



Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Word Embedding: word2vec

- “Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, **two-layer neural networks** that are trained to reconstruct linguistic contexts of words”
- “Word2vec **takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions**, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that **words that share common contexts in the corpus are located close to one another in the space**”

19

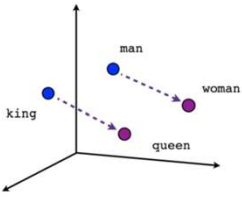
19

Explotación de la información, Clasificación y Agrupamiento de Información

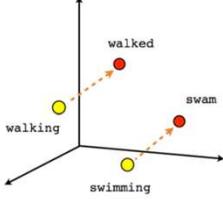
0. Introducción

Word Embedding: word2vec (cont.)

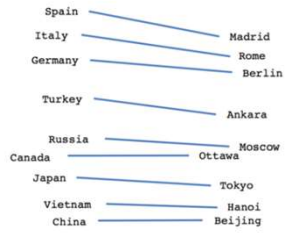
- Word vs. Sentence embeddings para sustituir el modelo bag of words de la RI tradicional:
 - <https://www.elastic.co/blog/text-similarity-search-with-vectors-in-elasticsearch>



Male-Female




Verb tense



Country-Capital

20



Explotación de la información, Clasificación y Agrupamiento de Información


Explotación de

0. Introducción

Word embeddings vs. Transformers (BERT):

- Transformers:
 - Generan vectores diferentes para un mismo término, dependiendo de en qué contexto aparece éste.
 - "banco", que puede ser el de sentarse, el de peces o el de guardar el dinero
 - Word2vec representaría estas tres acepciones con el mismo vector, mientras que BERT asignaría diferentes representaciones dependiendo del contexto (es decir, de las palabras que lo rodean).
 - Generan el modelo original entrenado y permiten tunearlo ("fine-tuning") para la tarea requerida (WSD, question answering, entity recognition, clasificación de textos...):
 - Fine-tuning de RoBERTa para clasificación de textos con 16 millones de tweets: 5 días de proceso
 - Siempre genera representaciones para cualquier palabra, aunque no la haya visto en el entrenamiento (trabaja a nivel de subpalabra):
 - Si Word2vec recibe una palabra que no ha visto durante el entrenamiento no le asignará ninguna representación.

21



Clasificación y Agrupamiento de Información

Explotación de

0. Introducción

Ex.Inf. Modulo 4. BERT de Google.pdf:


- BERT: "*Bidirectional Encoder Representations from Transformer*"
- "Cómo funciona BERT, la inteligencia artificial con la que Google quiere conseguir que su motor de búsqueda nos entienda mejor"

El xxx es el tipo de vegetación dominante en muchos xxx, incluidas las marismas salinas, pantanos y estepas. Los biomas dominados por pastos se XXX praderas. Estos biomas cubren el 31% del XXX de la Tierra. Los pastos son buenos para muchos XXX que pastan, como el ganado, los ciervos y los pequeños roedores como los ratones y los ratones de campo.

texto para predecir que faltan "pasto", "hábitats", "llaman", "suelo" y "mamíferos".

22

22




Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Transformers:

- OpenAI, el generador de textos fake creíbles (GPT-3):
 - A partir de libros públicos, toda la Wikipedia y millones de páginas web y documentos científicos disponibles en Internet
 - Es un modelo de lenguaje:
 - # Su objetivo es predecir qué es lo siguiente que viene en función de los datos previos. Es como una especie de “autocompletado”.
 - # Puedes por ejemplo escribir dos o tres frases de un artículo y GPT-3 se encargará de escribir el resto del artículo. También puedes generar conversaciones y las respuestas estarán basadas en el contexto de las preguntas y respuestas anteriores.


23




Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Transformers (cont.):

- Librería Huggingface  :
 - <https://huggingface.co/transformers/>
 - “...provides general-purpose architectures (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet...) for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over 32+ pretrained models in 100+ languages and deep interoperability between Jax, PyTorch and TensorFlow ...”
 - Repositorio con miles de modelos preentrenados: <https://huggingface.co/models>

24



Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Transformers. Librería Huggingface (cont.):

- Aplicaciones
 - (https://huggingface.co/transformers/task_summary.html):
 - Sequence Classification
 - Extractive Question Answering
 - Language Modeling
 - Named Entity Recognition
 - Summarization
 - Translation

25



Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Transformers. Librería Huggingface (cont.):

- Aplicaciones:
 - Sequence Classification

```

>>> from transformers import pipeline
>>> classifier = pipeline("sentiment-analysis")
>>> result = classifier("I hate you")[0]
>>> print(f"label: {result['label']}, with score: {round(result['score'], 4)}")
label: NEGATIVE, with score: 0.9991

>>> result = classifier("I love you")[0]
>>> print(f"label: {result['label']}, with score: {round(result['score'], 4)}")
label: POSITIVE, with score: 0.9999

```

- Extractive Question Answering
- Language Modeling
- Named Entity Recognition
- Summarization
- Translation

26

0. Introducción

Transformers. Librería Huggingface (cont.):

- Aplicaciones:
 - Sequence Classification
 - Extractive Question Answering

```

>>> from transformers import pipeline

>>> question_answerer = pipeline("question-answering")

>>> context = r"""
... Extractive Question Answering is the task of extracting an answer from a text given a question. An example of a
... question answering dataset is the SQuAD dataset, which is entirely based on that task. If you would like to fine-tune
... a model on a SQuAD task, you may leverage the examples/pytorch/question-answering/run_squad.py script.
... """

>>> result = question_answerer(question="What is extractive question answering?", context=context)
>>> print(f"Answer: '{result['answer']}', score: {round(result['score'], 4)}, start: {result['start']}, end: {result['end']}")
Answer: 'the task of extracting an answer from a text given a question', score: 0.6177, start: 34, end: 95

>>> result = question_answerer(question="What is a good example of a question answering dataset?", context=context)
>>> print(f"Answer: '{result['answer']}', score: {round(result['score'], 4)}, start: {result['start']}, end: {result['end']}")
Answer: 'SQuAD dataset', score: 0.5152, start: 147, end: 160

```

- Summarization
- Translation

27

0. Introducción

Transformers. Librería Huggingface (cont.):

- Language Modeling

```

>>> from transformers import pipeline
>>> unmasker = pipeline("fill-mask")

>>> from pprint import pprint
>>> pprint(unmasker(f"HuggingFace is creating a {unmasker.tokenizer.mask_token} that the community uses to solve NLP tasks."))
[{'score': 0.1793,
  'sequence': 'HuggingFace is creating a tool that the community uses to solve '
              'NLP tasks.',
  'token': 3944,
  'token_str': ' tool'},
 {'score': 0.1135,
  'sequence': 'HuggingFace is creating a framework that the community uses to '
              'solve NLP tasks.',
  'token': 7208,
  'token_str': ' framework'},
 {'score': 0.0524,
  'sequence': 'HuggingFace is creating a library that the community uses to '
              'solve NLP tasks.',
  'token': 5560,
  'token_str': ' library'},
 {'score': 0.0349,
  'sequence': 'HuggingFace is creating a database that the community uses to '
              'solve NLP tasks.',
  'token': 8503,
  'token_str': ' database'},
 {'score': 0.0286,
  'sequence': 'HuggingFace is creating a prototype that the community uses to '
              'solve NLP tasks.',
  'token': 17715,
  'token_str': ' prototype'}]

```

28

Explotación de la información. Clasificación y Agrupamiento de Información

0. Introducción

Transformers. Librería Huggingface 🤖 (cont.):

- Aplicaciones:
 - Text Generation

```
>>> from transformers import pipeline

>>> text_generator = pipeline("text-generation")
>>> print(text_generator("As far as I am concerned, I will", max_length=50, do_sample=False))
[{'generated_text': 'As far as I am concerned, I will be the first to admit that I am not a fan of the idea of a "free market." I think that the idea of a free market is a bit of a stretch. I think that the idea'}]
```

29

Explotación de la información. Clasificación y Agrupamiento de Información

0. Introducción

Transformers. Librería Huggingface 🤖 (cont.):

- Aplicaciones:
 - Named Entity Recognition

```
>>> from transformers import pipeline

>>> ner_pipe = pipeline("ner")

>>> sequence = """Hugging Face Inc. is a company based in New York City. Its headquarters are in DUMBO,
... therefore very close to the Manhattan Bridge which is visible from the window."""

>>> for entity in ner_pipe(sequence):
...     print(entity)
...
{'entity': 'I-ORG', 'score': 0.9996, 'index': 1, 'word': 'Hu', 'start': 0, 'end': 2}
{'entity': 'I-ORG', 'score': 0.9910, 'index': 2, 'word': '##gging', 'start': 2, 'end': 7}
{'entity': 'I-ORG', 'score': 0.9982, 'index': 3, 'word': 'Face', 'start': 8, 'end': 12}
{'entity': 'I-ORG', 'score': 0.9995, 'index': 4, 'word': 'Inc', 'start': 13, 'end': 16}
{'entity': 'I-LOC', 'score': 0.9994, 'index': 11, 'word': 'New', 'start': 40, 'end': 43}
{'entity': 'I-LOC', 'score': 0.9993, 'index': 12, 'word': 'York', 'start': 44, 'end': 48}
{'entity': 'I-LOC', 'score': 0.9994, 'index': 13, 'word': 'City', 'start': 49, 'end': 53}
{'entity': 'I-LOC', 'score': 0.9863, 'index': 19, 'word': 'D', 'start': 79, 'end': 80}
{'entity': 'I-LOC', 'score': 0.9514, 'index': 20, 'word': '##UM', 'start': 80, 'end': 82}
{'entity': 'I-LOC', 'score': 0.9337, 'index': 21, 'word': '##BO', 'start': 82, 'end': 84}
{'entity': 'I-LOC', 'score': 0.9762, 'index': 28, 'word': 'Manhattan', 'start': 114, 'end': 123}
{'entity': 'I-LOC', 'score': 0.9915, 'index': 29, 'word': 'Bridge', 'start': 124, 'end': 130}
```

30

Exploitación de la información. Clasificación y Agrupamiento de Información

0. Introducción

Transformers. Librería Huggingface (cont.):

Summariza:

```
>>> print(summarizer(ARTICLE, max_length=130, min_length=30, do_sample=False))
[{'summary_text': ' Liana Barrientos, 39, is charged with two counts of "offering a false instrument for filing in the first degree" In total, she has been married 10 times, with nine of her marriages occurring between 1999 and 2002 . At one time, she was married to eight men at once, prosecutors say .'}]]
```

```
>>> from transformers import pipeline
>>> summarizer = pipeline("summarization")

>>> ARTICLE = """ New York (CNN)When Liana Barrientos was 23 years old, she got married in Westchester County, New York.
... A year later, she got married again in Westchester County, but to a different man and without divorcing her first husband.
... Only 18 days after that marriage, she got hitched yet again. Then, Barrientos declared "I do" five more times, sometimes only within two weeks of each other.
... In 2010, she married once more, this time in the Bronx. In an application for a marriage license, she stated it was her "first and only" marriage.
... Barrientos, now 39, is facing two criminal counts of "offering a false instrument for filing in the first degree," referring to her false statements on the
... 2010 marriage license application, according to court documents.
... Prosecutors said the marriages were part of an immigration scam.
... On Friday, she pleaded not guilty at State Supreme Court in the Bronx, according to her attorney, Christopher Wright, who declined to comment further.
... After leaving court, Barrientos was arrested and charged with theft of service and criminal trespass for allegedly sneaking into the New York subway through an
... Annette Markowski, a police spokeswoman. In total, Barrientos has been married 10 times, with nine of her marriages occurring between 1999 and 2002.
... All occurred either in Westchester County, Long Island, New Jersey or the Bronx. She is believed to still be married to four men, and at one time, she was married
... Prosecutors said the immigration scam involved some of her husbands, who filed for permanent residence status shortly after the marriages.
... Any divorces happened only after such filings were approved. It was unclear whether any of the men will be prosecuted.
... The case was referred to the Bronx District Attorney's Office by Immigration and Customs Enforcement and the Department of Homeland Security's
... Investigation Division. Seven of the men are from so-called "red-flagged" countries, including Egypt, Turkey, Georgia, Pakistan and Mali.
... Her eighth husband, Rashid Rajput, was deported in 2006 to his native Pakistan after an investigation by the Joint Terrorism Task Force.
... If convicted, Barrientos faces up to four years in prison. Her next court appearance is scheduled for May 18.
... """
```

31

Exploitación de la información. Clasificación y Agrupamiento de Información

0. Introducción

Transformers. Librería Huggingface (cont.):

Aplicaciones:

- Translation

```
>>> from transformers import pipeline

>>> translator = pipeline("translation_en_to_de")
>>> print(translator("Hugging Face is a technology company based in New York and Paris", max_length=40))
[{'translation_text': 'Hugging Face ist ein Technologieunternehmen mit Sitz in New York und Paris.'}]
```

32

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

T5 Transformer de Google:

- Base del nuevo modelo del buscador (MUM) que sustituye a BERT:
- 1.000 veces más poderoso que BERT: multimodal y multilingüe

Diagram of our text-to-text framework. Every task we consider uses text as input to the model, which is trained to generate some target text. This allows us to use the same model, loss function, and hyperparameters across our diverse set of tasks including translation (green), linguistic acceptability (red), sentence similarity (yellow), and document summarization (blue). It also provides a standard testbed for the methods included in our empirical survey.

33

Explotación de la información, Clasificación y Agrupamiento de Información

0. Introducción

Ex.Inf. Modulo 4. Los algoritmos lo saben todo o deben ayudarles los humanos.pdf:

- Problema de no poder explicar porqué toman las decisiones que toman
- Problema del conjunto de datos de los que aprenden:
 - ¿Son equilibrados?
 - ¿Son parciales?
 - "How I'm fighting bias in algorithms":
 - https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms
 - MIT grad student Joy Buolamwini was working with facial analysis software when she noticed a problem: the software didn't detect her face -- because the people who coded **the algorithm hadn't taught it to identify a broad range of skin tones and facial structures.**

34

Explotación de la información, Clasificación y Agrupamiento de Información

1. Sistemas de clasificación de Información

Clasificación automática (*automated classification*):

- Asignación de una categoría predefinida disjunta
 - Distinto del proceso de **categorización** (*categorization*):
 - # Se permite la asignación de más de una clase, etiqueta o categoría para cada instancia: p.ej. asignar temáticas a libros
- Resultado: ontologías, taxonomías, jerarquías, vocabularios controlados o tesauros
- Problemas: precisión, consistencia, etc.
- Aplicaciones:
 - RI como un problema de clasificación con las categorías documento relevante/no relevante
 - Detección de spam o detección de páginas con contenido violento
 - Detección de autor (*Authorship attribution*)
 - *Part of speech tagging*
 - *Fluency ranking* en generación de texto

35

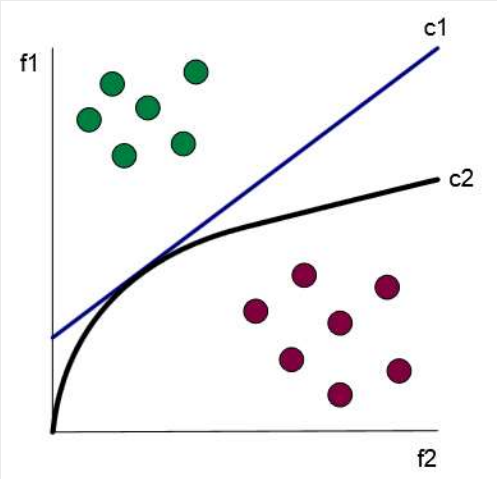
35

Explotación de la información, Clasificación y Agrupamiento de Información

1. Sistemas de clasificación de Información

■ **Objetivo:**

- Características f_1 y f_2
- Dos clases:
 - # Maximizar separación entre las clases



36

36

Explotación de la información, Clasificación y Agrupamiento de Información

1. Sistemas de clasificación de Información

Técnicas:

- Técnicas basadas en vocabulario:
- Árboles de decisión
- Basadas en reglas
- Estadísticas: co-ocurrencia de términos, redes neuronales, etc.

37

37

Explotación de la información, Clasificación y Agrupamiento de Información

1. Sistemas de clasificación de Información

Técnicas:

- Técnicas basadas en vocabulario:
 - Usan un tesoro o diccionario para determinar aquellos términos y sus variantes asociados a cada categoría
 - Problema: ambigüedad del lenguaje
- Árboles de decisión
- Basadas en reglas
- Estadísticas: co-ocurrencia de términos, redes neuronales, etc.

38

38

Explotación de la información, Clasificación y Agrupamiento de Información

2. Sist. clasificación información basadas en vocabulario

■ Ej.: categorías de comida

```

graph TD
    BG["(Broader)  
baked goods"] --> C["(Preferred)  
crackers"]
    C --> MT["(Narrower)  
melba toast"]
    C --> CH["(Related)  
cheese"]
    C --> PP["(Related)  
party planning"]
    B["(Variant)  
biscuits"] --> C
    CR["(Variant)  
crisps"] --> C
  
```

Example
Personal Digital Assistant

Synonyms
Handheld Computer

"Alternate" Spellings
Persenal Digitel Asistent

Abbreviations / Acronyms
PDA

Broader Terms
Wireless, Computers

Narrower Terms
PalmPilot, PocketPC

Related Terms
WindowsCE, Cell Phones

39

39

Explotación de la información, Clasificación y Agrupamiento de Información

2. Sist. clasificación información basadas en vocabulario

Ejercicio 1:

■ Obtener las reglas y vocabulario para las siguientes categorías de tipo de pregunta de los sistemas de búsqueda de respuesta. Utilizad a modo de ejemplo las preguntas que aparecen en la siguiente transparencia:

■ En la siguiente URL se pueden encontrar ayudas de sinónimos y relaciones semánticas:

<http://adimen.si.chu.es/cgi-bin/wei/public/wei.consult.perl>

Word: Look up

Word: Nouns: English_3.0:

near_synonym: English_3.0:

☒ Gloss ☒ English_3.0 ☐ Catalan_3.0
☐ Score ☐ Basque_3.0 ☐ Portuguese_3.0
☐ Rels ☐ Spanish_3.0
☐ Full ☐ Galician_3.0

40

40



2. Sist. clasificación información basadas en vocabulario


Explotación de la información, Clasificación y Agrupamiento de Información

entidad persona	¿como se llama el hijo de kim il sung?
entidad persona	¿quien es el creador de "doctor snuggles"?
entidad persona	¿quien es el lider bosnio?
entidad persona	¿quien fue la ganadora del torneo de wimbledon?
entidad persona	¿que presidente de corea del norte murio a los 82 años de edad?
entidad persona	¿quien es el presidente del parlamento europeo?
entidad persona	¿quien es el lider del sinn fein?
entidad persona	¿quien es el mayor exportador europeo de aceite de oliva?
entidad persona	¿quien escribio "star trek"?
entidad persona	¿quien es el presidente de la republica de italia?
entidad persona	¿quien ostenta el poder en pyongyang?
entidad persona	¿quien dirigio "con la muerte en los talones"?
entidad persona	¿quien es el presidente de rusia?
entidad persona	¿quien es el presidente italiano de asuntos exteriores?
entidad persona	¿quien es el entrenador del equipo nacional de futbol noruego?
entidad persona	¿quien es el director de la cia?
entidad persona	¿como se llamaba el cantante y lider de nirvana?
entidad persona	¿quien es el presidente de la republica francesa?
entidad persona	¿que primer ministro frances se suicido en los años 90?
entidad persona	¿quien es el presidente de peru?
entidad persona	¿que presidente ruso asistio a la reunion del g7 en napoles?
entidad persona	¿a que primer ministro abrio la fiscalia de milan un sumario por corrupcion?
entidad persona	¿quien proyecto la construccion de la catedral de san pedro?
entidad persona	¿como se llama el jefe de gobierno de australia?
entidad persona	¿como se llama el sucesor del gatt?
entidad persona	¿quien es el presidente de yugoslavia?
entidad persona	¿que ciudadano britanico recibio 50 latigazos en qatar?

entidad abreviat	¿cuales son las siglas del fondo mundial para la proteccion de la naturaleza?
entidad abreviat	¿cual es el acronimo de amnistia internacional?

41

41




2. Sist. clasificación información basadas en vocabulario

Explotación de la información, Clasificación y Agrupamiento de Información

entidad objeto	¿cual es la anterior moneda argentina?
entidad objeto	¿de que obtendra microsoft la licencia de sun?
entidad objeto	¿cual es el nombre de estandar europeo de comunicaciones moviles digitales?
entidad objeto	¿que produce la compañía victorinox?
entidad objeto	¿que produce mico?
entidad objeto	¿cual es el simbolo de paris?
entidad objeto	¿que tecnologia produce leica?
entidad objeto	¿como se llama el ferry naufragado en suecia en 1994?
entidad objeto	¿contra que choco el titanic?
entidad objeto	¿cual es el simbolo de liderazgo del giro de italia?
entidad objeto	¿que fue levantado el 13 de agosto de 1961?
entidad objeto	¿cual es la moneda iraki?
entidad objeto	nombre un edificio envuelto por christo.
entidad objeto	nombre una pelicula en la que se hayan usado animaciones por ordenador.
entidad objeto	¿que deporte practica adrian mutu?
entidad objeto	¿que alfabeto tiene solo cuatro letras "a, c, g, y t"?
entidad objeto	¿que plataforma estaba acampada en el paseo de la castellana de madrid?
entidad objeto	¿a que enfermedad corresponden las siglas rsi?
entidad objeto	¿que tipo de dolencia es caracteristica del rsi?
entidad objeto	¿que vitaminas ayudan en la lucha contra el cancer?
entidad objeto	¿que fruta tiene vitamina c?
entidad objeto	¿con el nombre de que enfermedad se corresponde el acronimo bse?
entidad objeto	¿que submarino choco con un buque en el canal de la mancha el 16 de febrero de 1995?
entidad objeto	¿en que epoca del año desaparecio jurgen schneider al producirse la bancarrota de su empres?
entidad objeto	¿que premio gano pulp fiction en el festival de cine de cannes?
entidad objeto	¿que nuevo canal de television gay aparecio en francia el 25 de octubre de 2004?
entidad objeto	¿cual es la ultima letra del alfabeto fonetico de la otan?
entidad objeto	¿con que pelicula marlee matlin gano un oscar?
entidad objeto	¿que huracan azoto la isla de cozumel?
entidad objeto	nombre una pelicula en la que haya participado kirk douglas en el periodo de 1946 a 1960.
entidad objeto	de el nombre de alguien que haya ganado el premio nobel de literatura entre 1945 y 1990.
entidad objeto	¿con que planeta choco el cometa shoemaker-levy?
entidad objeto	¿cual es la palabra alemana mas larga?
entidad objeto	¿como se llama la moneda de letonia?
entidad objeto	¿en que calle vive el primer ministro britanico?

42

42



2. Sist. clasificación información basadas en vocabulario


numerico econo	¿cuanto valen 10 pesos?
numerico econo	¿cuanto costo el túnel del canal?
numerico econo	¿que gasto se ha programado en virtud del ítop en el periodo 1994-1999 para la renovacion de
numerico econo	¿a cuanto ascendieron los beneficios del grupo fines de electronica y comunicaciones nokia e
numerico econo	¿cuanto reclama el sevilla fc a diego maradona?
numerico econo	¿a cuanto asciende el premio para la ganadora de wimbledon?
numerico econo	¿cual era el valor aproximado de la carga de un galeon del siglo xvi?
numerico econo	¿cuanto dinero gana anualmente el narcotrafico?
numerico econo	¿cual es el presupuesto de la interpol?
numerico econo	¿a cuanto asciende la multa que se le impuso a italia por superar la cuota de produccion de le
numerico econo	¿a cuanto ascendio la multa a john fashanu?

numerico medid	¿cual es la distancia entre la tierra y el sol?
numerico medid	¿que magnitud tuvo el terremoto que sacudio el norte de japon?
numerico medid	¿cuanto mide el everest?
numerico medid	¿que distancia se recorre en el rally granada-dakar?
numerico medid	¿cual es la extension de la selva lacandona?
numerico medid	¿cual es la distancia entre braga y guimarães?
numerico medid	¿cual es la altura del k2?
numerico medid	¿cual es la superficie de la baja sajonia?
numerico medid	¿a que distancia de burgos esta atapuerca?
numerico medid	¿a que distancia de la tierra esta jupiter?
numerico medid	¿que altura tiene el kanchenjunga?
numerico medid	¿que altura tiene la torre eiffel?
numerico medid	¿cuantos kilometros se recorrieron en el tour de 1926?
numerico medid	¿cual es el record del mundo de salto de altura?

numerico edad	¿a que edad murio joseph di mambro?
numerico edad	¿a que edad murio thomas "tip" o'neill?
numerico edad	¿cual era la esperanza de vida en francia en 1991?
numerico edad	¿que edad tenia nick leeson en el momento de ser condenado a la carcel?
numerico edad	¿que edad tenia richard holbrooke en 1995?

Explotación de la información, Clasificación y Agrupamiento de Información

43




3. Clasificación utilizando árboles de decisión

Técnicas de clasificación basadas en árboles de decisión:

- Construyen un modelo, hipótesis o representación de la regularidad existente en los datos
- Ventajas respecto a las redes neuronales o las máquinas de vectores de soporte (*Support Vector Machine, SVM*):
 - Son modelos comprensibles porque se pueden expresar de una manera simbólica, en forma de conjunto de condiciones
 - Son eficientes por su característica de algoritmos “voraces” (siempre que quepan todos los ejemplos para aprender en memoria)
 - Hay múltiples implementaciones disponibles
- Desventajas:
 - Son más dependientes del conjunto de ejemplos de aprendizaje

Explotación de la información, Clasificación y Agrupamiento de Información

44



Explotación de la información, Clasificación y Agrupamiento de Información

45

3. Clasificación utilizando árboles de decisión

Árbol de decisión:

- Conjunto de condiciones exhaustivas y excluyentes organizadas en una estructura jerárquica
 - Exhaustivo: cada condición ha de cumplirse una de sus opciones (edad > 50 ó edad ≤ 50)
 - Excluyente: las particiones del árbol han de ser disjuntas
- La decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas

45

45



Explotación de la información, Clasificación y Agrupamiento de Información

46

3. Clasificación utilizando árboles de decisión

Ejemplo:

Elección de “Play”

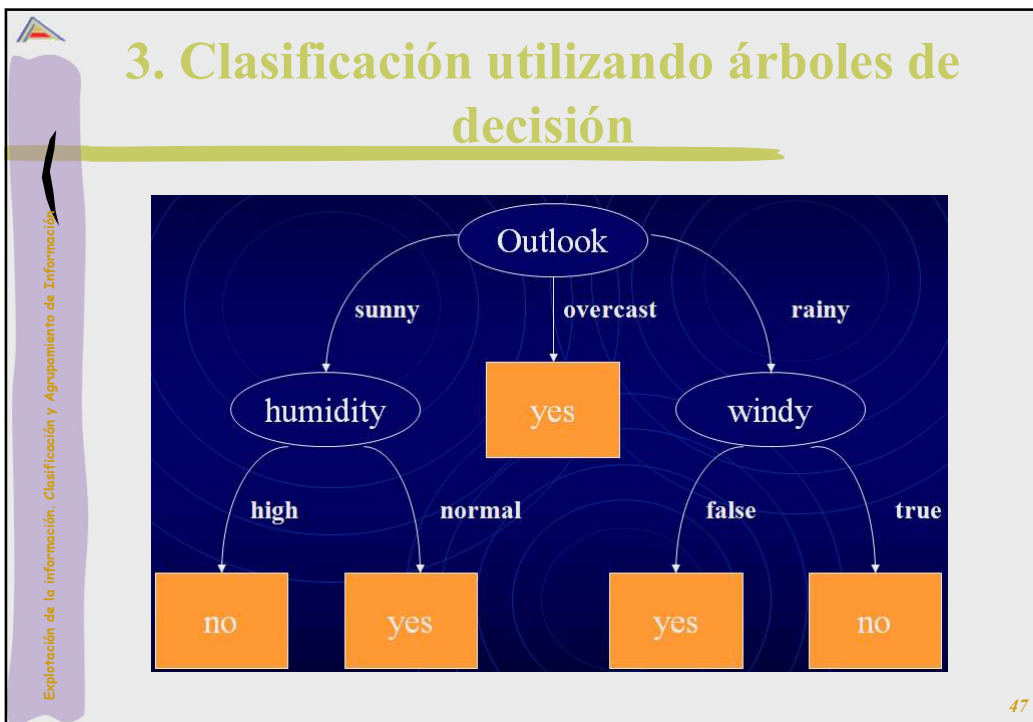
<http://csie.org/~dm/>

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

46

46

23



47

Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Algoritmo para construir AD a partir de datos:

- Técnica de partición (divide y vencerás):
 - El espacio de instancias se va partiendo de arriba abajo utilizando cada vez una partición o conjunto de condiciones excluyentes y exhaustivas
 - Una vez elegida la partición, dicha partición no se puede cambiar: *criterio de partición*

48

48

Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

Algoritmo partición (N:nodo; E:conj_ejemplos)

Si todos los ejemplos E son de la misma clase c

Entonces

 Asignar clase c al nodo N

 Salir // N es hoja

Sino

 particiones = generarPosiblesParticiones

 MejorPartición = seleccionarMejorParticiónSegún_criterio_partición

 Para cada condición i de la partición elegida

 Añadir un nodo hijo i a N y asignar los ejemplos consistentes (E_i)

partición (i, E_i) // Llamada recursiva

49

49

Explotación de la información, Clasificación y Agrupamiento de Información

3. Clasificación utilizando árboles de decisión

generarPosiblesParticiones:

■ Tipos de particiones:

- Nominales (x_i): aquellos que tienen un conjunto de posibles valores $\{v_1, v_2, \dots, v_k\}$
 - # Si solo se permiten árboles binarios, la partición será :
 - $(x_i=v_1, x_i \neq v_1), (x_i=v_2, x_i \neq v_2), (x_i=v_3, x_i \neq v_3), \dots$
 - # Caso contrario: $(x_i=v_1, x_i=v_2, \dots, x_i=v_k)$
- Numéricas (x_i): aquellos que tienen un conjunto de posibles valores numéricos y continuos. Las particiones: $(x_i \leq a, x_i > a)$, con a una constante numérica elegida entre un conjunto finito de constantes obtenidas de los ejemplos:
 - # Si x_i presenta los valores $\{0,2 \ 0,3 \ 0,7 \ 0,1 \ 0,8 \ 0,45 \ 0,33 \ 0,1 \ 0,8 \ 0\}$
 - # Se ordenan, eliminan repetidos $\{0 \ 0,1 \ 0,2 \ 0,3 \ 0,33 \ 0,45 \ 0,7 \ 0,8\}$ y se obtienen los valores intermedios $\{0,05 \ 0,15 \ 0,25 \ 0,315 \ 0,39 \ 0,575 \ 0,75\}$ generando particiones binarias:
 - $(x_i \leq 0,05, x_i > 0,05) (x_i \leq 0,15, x_i > 0,15) (x_i \leq 0,25, x_i > 0,25) (x_i \leq 0,315, x_i > 0,315) (x_i \leq 0,39, x_i > 0,39) (x_i \leq 0,575, x_i > 0,575) (x_i \leq 0,75, x_i > 0,75)$

50

50

3. Clasificación utilizando árboles de decisión

Ejercicio 2:

- Sobre el ejemplo anterior de elección de “Play”, a partir de la tabla de ejemplos, obtener las particiones
- Para n atributos y m valores posibles para cada atributo, ¿cuántas particiones se generarían?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

51

3. Clasificación utilizando árboles de decisión

seleccionarMejorParticiónSegún_criterio_partición:

- Objetivo: buscar particiones que discriminen más
- Criterio: elegir la partición s con mayor valor $I(s)$

$$I(s) = \sum_{j=1..n} p_j \cdot f(p_j^1, p_j^2, \dots, p_j^c)$$

- n : número de nodos hijos de la partición
- p_j : probabilidad de caer en el nodo j de la partición s
- p_j^l : proporción de elementos de la clase l en el nodo j
- c : número de clases del problema

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

52

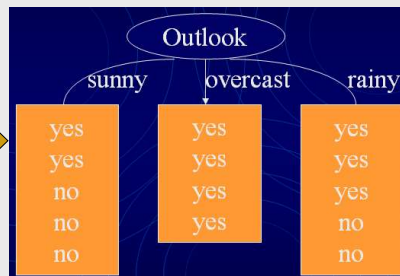
3. Clasificación utilizando árboles de decisión

Método basado en entropía C4.5 [Quinlan,93]:

$$I(s) = \sum_{j=1..n} p_j \times f(p_j^1, p_j^2, \dots, p_j^c) = \sum_{j=1..n} \left(p_j \times \sum_{k=1..c} (p_j^k \times \log_2(p_j^k)) \right)$$

$$I(Outlook) = \left(\frac{5}{14}\right) \times \left(\frac{2}{5} \times \log_2 \frac{2}{5} + \frac{3}{5} \times \log_2 \frac{3}{5}\right) + \left(\frac{4}{14}\right) \times 0 + \left(\frac{5}{14}\right) \times (-0.971) = -0.693$$

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



53

53

3. Clasificación utilizando árboles de decisión

Ejercicio 3:

- Sobre el ejemplo anterior de predicción del tiempo, calcula $I(s)$ para el resto de particiones
- ¿Qué partición quedaría como raíz del árbol de decisión final?

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

54

54

3. Clasificación utilizando árboles de decisión

Ejercicio 4:

■ Aplica el algoritmo *partición* para comprobar que se genera el árbol de decisión mostrado anteriormente

55

55

3. Clasificación utilizando árboles de decisión

Ejercicio 5: Calcula el árbol de decisión

Attributes					Class
Education	Annual Income	Age	Own House	Sex	Credit ranking
College	High	Old	Yes	Male	Good
High school	-----	Middle	Yes	Male	Good
High school	Middle	Young	No	Female	Good
College	High	Old	Yes	Male	Poor
College	High	Old	Yes	Male	Good
College	Middle	Young	No	Female	Good
High school	High	Old	Yes	Male	Poor
College	Middle	Middle	-----	Female	Good
High school	Middle	Young	No	Male	Poor

56

56

Explotación de la información, Clasificación y Agrupamiento de Información

4. Clasificación utilizando sistemas de reglas

Sistemas de reglas:

- Generalización de los árboles de decisión en el que **no se exige exclusión ni exhaustividad** en las condiciones de las reglas:
 - Se podría aplicar más de una regla (reglas 1, 3, 5) o ninguna
 - Se agrupan diferentes ramas del árbol en una sola condición: “en otro caso”
- Algoritmo:
 - Se generan reglas sucesivamente, descartándose ejemplos ya cubiertos por las reglas ya obtenidas, y con los ejemplos que quedan se empieza de nuevo

57

57

Explotación de la información, Clasificación y Agrupamiento de Información

4. Clasificación utilizando sistemas de reglas

Sistema para determinar la recomendación de cirugía ocular:

```

graph TD
    A[¿Astigmatismo?] -- no --> B[¿Edad?]
    A -- sí --> C[¿Miopía?]
    B -- "<25" --> D1[NO]
    B -- ">25 y <50" --> E[¿Miopía?]
    B -- ">50" --> D2[NO]
    E -- "<1.5" --> D3[NO]
    E -- ">1.5 y <10" --> D4[SÍ]
    E -- ">10" --> D5[NO]
    C -- "<6" --> D6[SÍ]
    C -- ">6" --> D7[NO]
    
```

1. SI astig = sí Y miopía>6 ENTONCES no
2. SI 25<edad≤50 Y miopía≤6 ENTONCES sí
3. SI edad>50 ENTONCES no
4. SI edad≤25 ENTONCES no
5. SI miopía>10 ENTONCES no
6. EN OTRO CASO operación=sí

SI astig = no Y 25<edad≤50 Y 1.5<miopía ≤ 10 ENTONCES sí

SI astig = sí Y miopía ≤ 6 ENTONCES sí

EN OTRO CASO no

58

58

4. Clasificación utilizando sistemas de reglas

Algoritmo cobertura(Epos, Eneg: conj_ejemplos)

Reglas = \emptyset

Mientras Epos $\neq \emptyset$ Y NO ParadaReglas // **Aprender nueva regla**

NuevaRegla = \emptyset

Eneg_Act = Eneg

Mientras Eneg_Act $\neq \emptyset$ Y NO ParadaCondiciones // **Aprender nueva condición**

Cond = seleccionar una condición según criterio (*elimina muchos negativos*)

NuevaRegla = NuevaRegla \cup {Cond} // Añadimos la nueva condición a la regla

Eneg_Act = ejemplos negativos consistentes con NuevaRegla

Reglas = Reglas \cup {NuevaRegla}

Epos = Epos – Ejemplos cubiertos por NuevaRegla

Retorna Reglas

59


59

Table 1.1 The contact lens data.

age	spectacle prescription	astigmatism	tear production rate	recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

60

60



Explotación de la información, Clasificación y Agrupamiento de Información

4. Clasificación utilizando sistemas de reglas


Datos de prescripción de lentes *hard*:

age=young	2/8
age=pre-presbyopic	1/8
age=presbyopic	1/8
spectacle prescription=myope	3/12
spectacle prescription=hypermetrope	1/12
astigmatism=no	0/12
astigmatism=yes	4/12
tear production rate=reduced	0/12
tear production rate=normal	4/12

■ Añadimos regla:

- SI astigmatism = yes ENTONCES recommendation = hard

61



Explotación de la información, Clasificación y Agrupamiento de Información


4. Clasificación utilizando sistemas de reglas

■ Cogemos el resto de ejemplos para refinarla:

Table 4.8 Part of the contact lens data for which **astigmatism = yes.**

age	spectacle prescription	astigmatism	tear production rate	recommended lenses		
					age=young	2/4
					age=pre-presbyopic	1/4
					age=presbyopic	1/4
young	myope	yes	reduced	none	spectacle prescription=myope	3/6
young	myope	yes	normal	hard		
young	hypermetrope	yes	reduced	none		
young	hypermetrope	yes	normal	hard		
pre-presbyopic	myope	yes	reduced	none	spectacle prescription=hypermetrope	1/6
pre-presbyopic	myope	yes	normal	hard		
pre-presbyopic	hypermetrope	yes	reduced	none		
pre-presbyopic	hypermetrope	yes	normal	none		
presbyopic	myope	yes	reduced	none	tear production rate=reduced	0/6
presbyopic	myope	yes	normal	hard		
presbyopic	hypermetrope	yes	reduced	none		
presbyopic	hypermetrope	yes	normal	none	tear production rate=normal	4/6

62



Explotación de la información, Clasificación y Agrupamiento de Información

63

4. Clasificación utilizando sistemas de reglas

- Regla refinada:
 - SI astigmatism = yes
Y tear production rate = normal,
ENTONCES recommendation = hard

63

63



Explotación de la información, Clasificación y Agrupamiento de Información

64

4. Clasificación utilizando sistemas de reglas


- Seguimos refinando:

Table 4.9 Part of the contact lens data for which astigmatism = yes and tear production rate = normal.

age	spectacle prescription	astigmatism	tear production rate	recommended lenses
young	myope	yes	normal	hard
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	normal	none

64

64



Explotación de la información, Clasificación y Agrupamiento de Información

65

4. Clasificación utilizando sistemas de reglas


■ Seguimos refinando:

age=young	2/2
age=pre-presbyopic	1/2
age=presbyopic	1/2
spectacle prescription=myope	3/3
spectacle prescription=hypermetrope	1/3

■ Regla refinada:

- SI astigmatism = yes
Y tear production rate = normal
Y spectacle prescription = myope,
ENTONCES recommendation = hard

65



Explotación de la información, Clasificación y Agrupamiento de Información

66

Table 1.1 The contact lens data.

age	spectacle prescription	astigmatism	tear production rate	recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

66

Explotación de la información, Clasificación y Agrupamiento de Información

4. Clasificación utilizando sistemas de reglas

- # **Ejercicio 6:**
 - Genera la regla para prescripción de lentes *soft*
- # **Ejercicio 7:**
 - Genera el árbol de decisión según el algoritmo *partición* visto anteriormente
- # **Ejercicio 8:**
 - Genera las reglas del ejercicio de elección de “play” visto anteriormente

67

67

Explotación de la información, Clasificación y Agrupamiento de Información

5. Problema del *overfitting*. Sistemas de poda

- # **Sistemas de poda:**
 - Puede darse situaciones de *overfitting*:
 - Que el modelo aprendido se ajuste en exceso a los ejemplos conocidos y funcione mal para los nuevos ejemplos
 - Especialmente cuando los ejemplos con los que se aprende contienen “ruido”
 - Solución:
 - # Obtención de modelos más generales:
 - Eliminando condiciones de las ramas del árbol o de algunas reglas

68

68

5. Problema del *overfitting*. Sistemas de poda

Prepoda:

- Se realiza durante la construcción del árbol o conjunto de reglas
- Se determina el criterio de parada para seguir especializando una rama o regla:
 - N° de ejemplos por nodo, n° de excepciones respecto a la clase mayoritaria, etc.

Pospoda:

- Después de la construcción del árbol o conjunto de reglas
- Se eliminan nodos o reglas en sentido ascendente
- Es menos eficiente que la prepoda

Prepoda + pospoda:

- Algoritmo C4.5 con prepoda por cardinalidad y pospoda más sofisticada

69

69

6. Part of speech tagging

Objetivo de un *POS tagger*:

- A/AT similar/JJ resolution/NN passed/VBD in/IN the/AT Senate/NN by/IN a/AT vote/NN of/IN 29-5/CD ./.
- Desambiguar:
 - (Verbo) I wouldn't **trust** him.
 - (Nombre) He put money in the family **trust**

Técnicas:

- Basadas en frecuencia de aparición del *tag*.
- Basadas en n-gramas
- Modelos estocásticos

70

70

Explotación de la información, Clasificación y Agrupamiento de Información

6. Part of speech tagging

Basadas en frecuencia de aparición del tag:

- $P(t_i | w) = c(w, t_i) / (c(w, t_1) + \dots + c(w, t_k))$
 - $c(w, t_i)$ = número de veces que w/t_i aparece en el corpus
- Éxito: 91% para inglés
- Ejemplo:
 - heat :: noun/89, verb/5

71

71

Explotación de la información, Clasificación y Agrupamiento de Información


6. Part of speech tagging

Transformation-based learning:

- A simple rule-based part of speech tagger. Brill. 1992
- Método:
 1. Etiquetar cada token con el tag más frecuente
 2. Crear reglas que corrijan tags erróneos
 - old_tag new_tag NEXT-TAG tag
 - old_tag new_tag PREV-TAG tag
 - TO IN NEXT-TAG AT
 - NN VB PREV-TAG TO
 3. Contar cuántas correcciones con éxito y fracaso se realizan con cada regla
 4. Seleccionar la mejor regla que maximice: |éxito| - |fracaso|
 5. Si no se alcanza un umbral, ir al paso 2

72

72



Explotación de la información, Clasificación y Agrupamiento de Información

6. Part of speech tagging

Ejercicio 9:

- Sobre el texto etiquetado del ejercicio 2 del módulo 2, obtener reglas que resuelvan errores de etiquetado aplicando la técnica de *Transformation-based learning*.

 - Una descripción más detallada de las etiquetas léxicas se puede encontrar en la siguiente transparencia y en <http://www.scs.leeds.ac.uk/ccalas/tagsets/brown.html>

 - # 020 CD 020 020
 - # THE NP the the
 - # REBELS NNS rebel rebel

73

73



Explotación de la información, Clasificación y Agrupamiento de Información

6. Part of speech tagging

UPenn TreeBank II word tags:

• CC - Coordinating conjunction	• PRP\$ - Possessive pronoun
• CD - Cardinal number	• RB - Adverb
• DT - Determiner	• RBR - Adverb, comparative
• EX - Existential there	• RBS - Adverb, superlative
• FW - Foreign word	• RP - Particle
• IN - Preposition or subordinating conjunction	• SYM - Symbol
• JJ - Adjective	• TO - to
• JJR - Adjective, comparative	• UH - Interjection
• JJS - Adjective, superlative	• VB - Verb, base form
• LS - List item marker	• VBD - Verb, past tense
• MD - Modal	• VBG - Verb, gerund or present participle
• NN - Noun, singular or mass	• VBN - Verb, past participle
• NNS - Noun, plural	• VBP - Verb, non-3rd person singular present
• NNP - Proper noun, singular	• VBZ - Verb, 3rd person singular present
• NNPS - Proper noun, plural	• WDT - Wh-determiner
• PDT - Predeterminer	• WP - Wh-pronoun
• POS - Possessive ending	• WP\$ - Possessive wh-pronoun
• PRP - Personal pronoun	• WRB - Wh-adverb

74

74

6. Part of speech tagging

Basadas en n-gramas:

■ Corpus de 1000 palabras

■ 1000 uni-gramas

“vice” aparece 50 veces

$$\blacksquare p(\text{vice}) = 50/1000 = 0.05$$

“presidente” aparece 100 veces

$$\blacksquare p(\text{presidente}) = 100/1000 = 0.1$$

“vice presidente”

$$\blacksquare p(\text{vice presidente}) = 40/1000 = 0.04$$

$$\blacksquare p(\text{vice}) * p(\text{presidente}) = 0.05 \times 0.01 = 0.005$$

■ 999 bigramas:

Bigrama “vice – presidente” aparece 40 veces:

$$\blacksquare p(\text{vice-presidente}) = 40/999 = 0.04$$

■ Objetivo: las bigramas con mayor *Pointwise Mutual Information* (PMI) son las más probables

$$PMI(W1W2) = \log \frac{p(W1W2)}{p(W1)p(W2)} = \log \frac{p(\text{vice presidente})}{p(\text{vice})p(\text{presidente})} = \log \frac{0.04}{0.005} = 2.08$$

75

75

6. Part of speech tagging. Basadas en n-gramas

Modelos ocultos de Markov:

■ “La predicción del siguiente estado solo depende del estado actual”

$$p(w_n | w_{n-1}) = \frac{C(w_{n-1}, n)}{C(w_{n-1})}$$

■ Probabilidad de una oración utilizando bigramas:

$$p(w_{0..n}) = \prod_{i=0}^n p(w_i | w_{i-1})$$

76

76

6. Part of speech tagging. Basadas en n-gramas

Trigramas:

- La estimación de máxima verosimilitud del trigrama “of the king”:

$$P_{MLE}(\text{KING} \mid \text{OF THE}) = \frac{\text{count}(\text{OF THE KING})}{\sum_w \text{count}(\text{OF THE } w)} = \frac{\text{count}(\text{OF THE KING})}{\text{count}_{\text{hist}}(\text{OF THE})}$$

77

77

6. Part of speech tagging

Modelos estocásticos:

- Dada la secuencia de palabras de una oración:
 - $W = w_1, w_2, \dots, w_n$
- Asignar una secuencia de etiquetas:
 - $T = t_1, t_2, \dots, t_n$
- Objetivo:
 - Encontrar T que maximice $P(T|W) = P(W|T) P(T) / P(W) = \alpha P(W|T) P(T)$
- Forma de cálculo:
 - $P(T) = P(t_1) P(t_2 \mid t_1) P(t_3 \mid t_1, t_2) P(t_4 \mid t_1, t_2, t_3) \dots P(t_n \mid t_1, t_2, \dots, t_{n-1}) \approx P(t_1) P(t_2 \mid t_1) P(t_3 \mid t_2) \dots P(t_n \mid t_{n-1})$
 - # Utilizando *second order Markov model*: $P(t_i \mid t_{i-2}, t_{i-1})$
 - $P(W|T) = P(w_1 \mid t_1) P(w_2 \mid t_2) \dots P(w_n \mid t_n)$
 - $P(w_i \mid t_j) = C(w_i, t_j) / C(t_j)$

78

78

6. Part of speech tagging. Modelos estocásticos

Table 1: Statistics to be collected.

notation	counting the number of
C_n	all word tokens w
$C(w)$	occurrences of the word w
$C(w, t)$	occurrences of the word w tagged with t
$C(t)$	occurrences of the tag t
$C(t_1, t_2)$	occurrences of the tag bigram (t_1, t_2) , that is the tag t_1 followed by the tag t_2
$C(t_1, t_2, t_3)$	occurrences of the tag trigram (t_1, t_2, t_3) , that is the tag t_1 followed by t_2 followed by t_3
$C(w_1, t_1, t_2)$	occurrences of the wordtag-tag bigram (w_1, t_1, t_2) , that is the word w_1 tagged with t_1 followed by the tag t_2
$C_m(t)$	different word types tagged with tag t
$C_c(t)$	occurrences of capitalized words tagged with t
$C_m(w_{\text{end-}i}, t)$	different word types ending with the same i letters w and tagged with t

$$P(t_i) = \frac{C(t_i)}{C_n}$$

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

$$P(t_i|t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-1}, t_{i-2})}$$

$$P(t_i|w_{i-1}, t_{i-1}) = \frac{C(w_{i-1}, t_{i-1}, t_i)}{C(w_{i-1}, t_{i-1})}$$

$$P(w_i|t_i) = \frac{C(w_i, t_i)}{C(t_i)}$$

$$P(t_i|w_i) = \frac{C(w_i, t_i)}{C(w_i)}$$

79

79

6. Part of speech tagging. Modelos estocásticos

Para ampliar conocimientos:

- “Implementing an efficient part-of-speech tagger”. Johan Carlberger, Viggo Kann. 24th March 1999

- Google Books: Ngram Viewer

- <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>

English

Version 20120701

[total counts](#)

1-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_

ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d da db dc
 dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e ea eb ec
 ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew
 ex ey ez f fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp
 fq fr fs ft fu fv fw fx fy fz g ga gb gc gd ge gf gh gi gj
 gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h ha hb hc
 hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw
 hx hy hz i ia ib ic id ie if ig ih ii ij ik il im in io ip iq
 ir is it iu iv iw ix iy iz j ja jb jc jd je jf jg jh ji jk jl jm
 jn jo jp jq jr js jt ju jv jw jx jy jz k ka kb kc kd ke kf kg
 kh ki kj kl km kn ko kp kq kr ks kt ku kv kw kx ky kz L la lb
 lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw
 lx ly lz m ma mb mc md me mf mg mh mi mj mk ml mn mo
 mp mq mr ms mt mu mv mw mx my mz n na nb nc nd ne nf ng
 nh ni nj nk nl nm no np nq nr ns nt nu nv nw nx ny nz o oa
 ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot ou
 ov ow ox oy oz p pa pb pc pd pe pf pg ph pi pj pk pl pm
 pn po pp pq pr ps pt pu pv pw px py pz q qa qb qc qd qe
 qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx
 qy qz r ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr
 rs rt ru rv rw rx ry rz s sa sb sc sd se sf sg sh si sj sk
 sl sm sn so sp sq sr ss st su sv sw sx sy sz t ta tb tc td te
 tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u

3-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_

ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d da db dc
 dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e ea eb ec
 ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew
 ex ey ez f fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp
 fq fr fs ft fu fv fw fx fy fz g ga gb gc gd ge gf gh gi gj
 gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h ha hb hc
 hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw
 hx hy hz i ia ib ic id ie if ig ih ii ij ik il im in io ip iq
 ir is it iu iv iw ix iy iz j ja jb jc jd je jf jg jh ji jk jl jm
 jn jo jp jq jr js jt ju jv jw jx jy jz k ka kb kc kd ke kf kg
 kh ki kj kl km kn ko kp kq kr ks kt ku kv kw kx ky kz L la lb
 lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw
 lx ly lz m ma mb mc md me mf mg mh mi mj mk ml mn mo
 mp mq mr ms mt mu mv mw mx my mz n na nb nc nd ne nf ng
 nh ni nj nk nl nm no np nq nr ns nt nu nv nw nx ny nz o oa
 ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot ou
 ov ow ox oy oz p pa pb pc pd pe pf pg ph pi pj pk pl pm
 pn po pp pq pr ps pt pu pv pw px py pz q qa qb qc qd qe
 qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx
 qy qz r ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr
 rs rt ru rv rw rx ry rz s sa sb sc sd se sf sg sh si sj sk
 sl sm sn so sp sq sr ss st su sv sw sx sy sz t ta tb tc td te
 tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u

4-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_

ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d da db dc
 dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e ea eb ec
 ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew
 ex ey ez f fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp
 fq fr fs ft fu fv fw fx fy fz g ga gb gc gd ge gf gh gi gj
 gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h ha hb hc
 hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw
 hx hy hz i ia ib ic id ie if ig ih ii ij ik il im in io ip iq
 ir is it iu iv iw ix iy iz j ja jb jc jd je jf jg jh ji jk jl jm
 jn jo jp jq jr js jt ju jv jw jx jy jz k ka kb kc kd ke kf kg
 kh ki kj kl km kn ko kp kq kr ks kt ku kv kw kx ky kz L la lb
 lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw
 lx ly lz m ma mb mc md me mf mg mh mi mj mk ml mn mo
 mp mq mr ms mt mu mv mw mx my mz n na nb nc nd ne nf ng
 nh ni nj nk nl nm no np nq nr ns nt nu nv nw nx ny nz o oa
 ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot ou
 ov ow ox oy oz p pa pb pc pd pe pf pg ph pi pj pk pl pm
 pn po pp pq pr ps pt pu pv pw px py pz q qa qb qc qd qe
 qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx
 qy qz r ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr
 rs rt ru rv rw rx ry rz s sa sb sc sd se sf sg sh si sj sk
 sl sm sn so sp sq sr ss st su sv sw sx sy sz t ta tb tc td te
 tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u

5-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_

ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d da db dc
 dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e ea eb ec
 ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew
 ex ey ez f fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp
 fq fr fs ft fu fv fw fx fy fz g ga gb gc gd ge gf gh gi gj
 gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h ha hb hc
 hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw
 hx hy hz i ia ib ic id ie if ig ih ii ij ik il im in io ip iq
 ir is it iu iv iw ix iy iz j ja jb jc jd je jf jg jh ji jk jl jm
 jn jo jp jq jr js jt ju jv jw jx jy jz k ka kb kc kd ke kf kg
 kh ki kj kl km kn ko kp kq kr ks kt ku kv kw kx ky kz L la lb
 lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw
 lx ly lz m ma mb mc md me mf mg mh mi mj mk ml mn mo
 mp mq mr ms mt mu mv mw mx my mz n na nb nc nd ne nf ng
 nh ni nj nk nl nm no np nq nr ns nt nu nv nw nx ny nz o oa
 ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot ou
 ov ow ox oy oz p pa pb pc pd pe pf pg ph pi pj pk pl pm
 pn po pp pq pr ps pt pu pv pw px py pz q qa qb qc qd qe
 qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx
 qy qz r ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr
 rs rt ru rv rw rx ry rz s sa sb sc sd se sf sg sh si sj sk
 sl sm sn so sp sq sr ss st su sv sw sx sy sz t ta tb tc td te
 tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u

dependencies 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _

80

80

6. Part of speech tagging. Modelos estocásticos

Ejercicio 10:

■ Dadas las dos siguientes frases:

1. Secretariat/NNP is/VBZ expected/VBN to/TO **race**/VB tomorrow/NN
2. People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT **race**/NN for/IN outer/JJ space/NN

■ Y dadas las probabilidades de las bigramas:

- $P(\text{NN}|\text{TO}) = .021$ $P(\text{race}|\text{NN}) = .00041$
- $P(\text{VB}|\text{TO}) = .34$ $P(\text{race}|\text{VB}) = .00003$

■ Calcular la etiqueta más probable para “race” según el modelo estocástico **para la frase 1.**

81

81

6. Part of speech tagging. Modelos estocásticos

Añadiendo reglas:

■ Detección de nombres propios si la palabra empieza por mayúscula:

$$P_c(w, t) = \begin{cases} \gamma_1 & \text{if } t \text{ is not proper-noun tag and } w \text{ is capitalized,} \\ \gamma_2 & \text{if } t \text{ is proper-noun tag and } w \text{ is not capitalized,} \\ 1 & \text{otherwise.} \end{cases}$$

- $\gamma_1 = 0.028$ and $\gamma_2 = 0.044$
- En el caso de palabras desconocidas: $\gamma_1 = 0.020$ $\gamma_2 = 0.048$

$$T(w_{1..n}) = \arg \max_{t_{1..n}} \prod_{i=1}^n P_{int}(t_i | t_{i-2}, t_{i-1}) P(w_i | t_i) P_c(w_i, t_i).$$

82

82

6. Part of speech tagging. Modelos estocásticos

Etiquetando palabras desconocidas:

- Hay que estimar $P_m(w | t)$ en lugar de $P(w | t)$: éxito del 45.5% en etiquetado de palabras desc.

$$P_m(w|t) = \frac{C_m(t)}{\sum_{\tau \in \text{tag set}} C_m(\tau)}$$

- Se puede añadir frecuencias de terminaciones (L máximo de 5, éxito del 88.7%):

$$P_e(w|t) = \sum_{i=0}^L \alpha_i \cdot \frac{C(w_{\text{end-}i}, t)}{\sum_{\tau \in \text{tag set}} C(w_{\text{end-}i}, \tau)}$$

83

83


7. Sistemas de agrupamiento de Información

Agrupamiento (*clustering*):

- Separar en grupos basándose en las similitudes o relaciones existentes
- Diferencias con la clasificación automática:
 - Los grupos o categorías no están necesariamente predefinidos
 - Se pueden asignar uno o varios grupos
- Aplicaciones:
 - Recuperación de información: organizar los resultados
 - Facilitar la navegación por una colección de documentos
 - Creación de directorios Web (*Yahoo*)

84

84

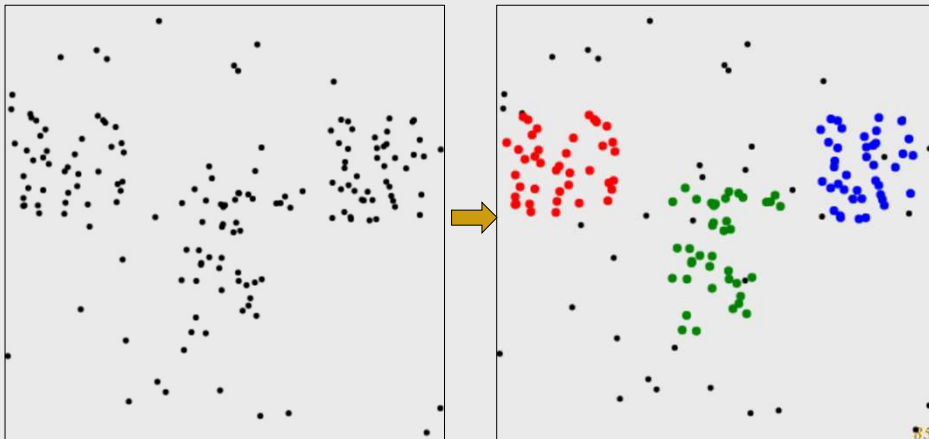


Explotación de la información, Clasificación y Agrupamiento de Información


7. Sistemas de agrupamiento de Información

Agrupamiento (*clustering*) (cont.)

https://www.w3schools.com/ai/ai_clustering.asp



85

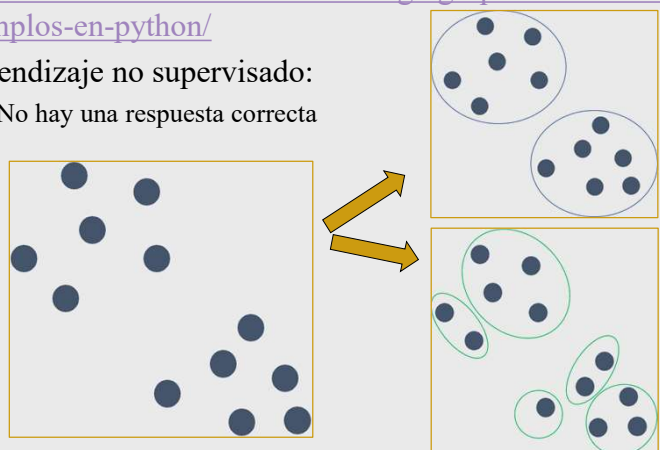


Explotación de la información, Clasificación y Agrupamiento de Información

7. Sistemas de agrupamiento de Información

Agrupamiento (*clustering*) (cont.):

- <https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/>
- Aprendizaje no supervisado:
 - # No hay una respuesta correcta



86

7. Sistemas de agrupamiento de Información

Agrupamiento en la RI:

■ Objetivo:

- Particionar una colección de documentos D en k subconjuntos o clusters D_1, D_2, \dots, D_k , de tal forma que se minimice la distancia intracluster o se maximice la semejanza intracluster:

✦ Utilizando el modelo vectorial:

- Un clúster sería un *centroide* de los documentos

■ Objetivo:

- Minimizar $\sum_i \sum_{d \in D_i} \text{distancia}(d, \bar{D}_i)$ o maximizar $\sum_i \sum_{d \in D_i} \text{semejanza}(d, \bar{D}_i)$

■ Hipótesis de agrupamiento:

- Los documentos fuertemente asociados tienden a ser relevantes para la misma consulta
- Si un usuario está interesado en un doc de un grupo, también es probable que lo esté en los demás miembros del grupo

87

87

8. Sistemas de agrupamiento de Información en la RI

Tipos de agrupamiento en la RI:

■ *Pre-retrieval document clustering*:


- Se realiza en fase de indexación
- Se elige un representante del grupo que sería con el que se compara la query (los restantes docs del grupo no se comparan)
- Problema: creación de grupos estáticos en un entorno tan dinámico como es la Web

■ *Post-retrieval document clustering*:

- Se realiza en fase de presentación de resultados de la fase de búsqueda
- Se agrupan los documentos devueltos por el motor de búsqueda
- Problema: eficiencia del proceso en tiempo de búsqueda

88

88



Explotación de la información, Clasificación y Agrupamiento de Información


8. Sistemas de agrupamiento de Información en la RI

Fases en el agrupamiento en la RI:

- Selección/extracción de características: representación de objetos
- Cálculo de la similitud entre objetos: medidas de distancia
- Clustering o agrupamiento

89

89



Explotación de la información, Clasificación y Agrupamiento de Información

8. Sistemas de agrupamiento de Información en la RI

Técnicas de agrupamiento:

- **No exclusivas:** un doc puede pertenecer a varios grupos
- **Exclusivas:** un doc solo pertenece a un grupo
 - **Extrínsecas:**
 - # Cuando los grupos están predefinidos y se tienen objetos que ya están agrupados en dichos clústeres, los cuales son utilizados por el algoritmo para aprender a agrupar el resto de objetos
 - **Intrínsecas:**
 - # Los grupos se crean a partir de las características propias de los objetos sin conocer previamente los grupos
 - # Tipos:
 - **Jerárquicas:** los grupos se consiguen mediante la separación o unión de grupos de documentos generando una estructura en árbol con grupos anidados
 - **Particionales:** se llega a un agrupamiento que optimiza un criterio predefinido o función objetivo, creando una estructura plana, sin grupos anidados

90

90

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Técnicas de agrupamiento particionales (*k*-clustering, *k*-means, *k*-medoids):

- Algoritmo:
 - Se determina a priori el *número de grupos*:
 - # Se cogen los *k* primeros objetos, o
 - # Los *k* objetos más alejados entre sí, o
 - # *k* objetos aleatoriamente
 - Iterativamente se van asignando docs a estas particiones
 - Los docs se reasignan de acuerdo a una *función objetivo*
 - El proceso se repite hasta que se consigue un *criterio de terminación*
- Variaciones de los clusters:
 - Juntar grupos cuando la distancia entre sus centroides esté por debajo de un umbral
 - Dividir grupos cuando su varianza esté por encima de un umbral

91

91

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Función objetivo:

- Internas: miden similitud *intra-cluster*:
 - Maximizar la suma de los promedios de las similitudes existentes entre los pares de docs asignados a cada clúster, teniendo en cuenta el tamaño de cada uno:
 - # *k*: n° de clústeres; *n*: n° elementos de cada clúster; *sim*(*d*, *e*): función de similitud p.ej. el coseno
$$\max I_1 = \sum_{r=1}^k n_r \times \left(\frac{1}{n_r^2} \times \sum_{d_i, d_j \in S_r} \text{sim}(d_i, d_j) \right)$$
- Externas: miden distancia *inter-cluster*:
 - Minimizar similitud entre centroide de cada clúster y el centroide de la colección completa
$$\min E_1 = \sum_{r=1}^k n_r \times \text{sim}(C_r, C)$$

92

92

9. Sistemas de agrupamiento de información particionales

k-mean:

- Generar los k clústeres iniciales con sus docs
- Inicializar los centroides de cada clúster
- Mientras sea posible realizar más mejoras
 - Para cada documento d
 - # Encontrar el clúster c cuyo centroide es más similar a d
 - # Asignar d al clúster c
 - Para cada clúster c
 - # Recalcular el centroide de c según los documentos asignados a c

93

93

9. Sistemas de agrupamiento de información particionales

Ejemplo de aplicación de *k-mean* (University of South Carolina Upstate, Angelina Tzacheva):

- Supongamos:
 - Los siguientes 8 vectores: $A_1(2, 10)$ $A_2(2, 5)$ $A_3(8, 4)$ $A_4(5, 8)$ $A_5(7, 5)$ $A_6(6, 4)$ $A_7(1, 2)$ $A_8(4, 9)$
 - $k=3$
 - Clusters iniciales: $A_1(2, 10)$, $A_4(5, 8)$, $A_7(1, 2)$
 - Distancia entre dos vectores $a=(x_1, y_1)$ y $b=(x_2, y_2)$:
 - # $\rho(a, b) = |x_2 - x_1| + |y_2 - y_1|$
 - Centroide de un grupo n de vectores: vector con el resultado de la media de los n vectores. Cada componente del vector centroide será la media aritmética de las casillas de todos los vectores

94

94

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Iteración 1 de *k-means*:

		Cluster 1 (2, 10)	Cluster 2 (5, 8)	Cluster 3 (1, 2)	
	Vector	Dist Clust 1	Dist Clust 2	Dist Clust 3	Cluster
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

95

95

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Iteración 1 de *k-means*:

▣ Clusters conseguidos:

Cluster 1 (2, 10)	Cluster 2 (8, 4) (5, 8) (7, 5) (6, 4) (4, 9)	Cluster 3 (2, 5) (1, 2)

96

96

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Iteración 1 de *k-means*:

- Recálculo de los centroides:
 - Cluster 1: (2, 10)
 - Cluster 2: $((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)$
 - Cluster 3: $((2+1)/2, (5+2)/2) = (1.5, 3.5)$

97

97

Explotación de la información, Clasificación y Agrupamiento de Información

9. Sistemas de agrupamiento de información particionales

Ejercicio 9:

- Calcular las dos siguientes iteraciones del algoritmo *k-means*

98

98

Explotación de la información, Clasificación y Agrupamiento de Información

10. Sistemas de agrupamiento de información jerárquicos

Dendograma:

- Estructura en árbol de clusters o grupos
- Las hojas son grupos que contienen un único documento
- La raíz es un único grupo con todos los elementos de la colección
- Los niveles intermedios son las posibles configuraciones de clusters

99

Explotación de la información, Clasificación y Agrupamiento de Información

10. Sistemas de agrupamiento de información jerárquicos

Tipos de sistemas jerárquicos:

- Aglomerativos:
 - Se comienza con los objetos o individuos de modo individual
 - Luego se van agrupando de modo que los primeros en hacerlo son los más similares
 - Al final, todos los subgrupos se unen en un único cluster
- Divisivos:
 - Se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén

100

10. Sistemas de agrupamiento de información jerárquicos

Sistemas jerárquicos aglomerativos. Algoritmo:

- Empezar con N clusters (el número inicial de elementos) y una matriz $N \times N$ simétrica de distancias o similitudes. $D = [d_{ik}]_{ik}$.
- Dentro de D, buscar aquella entre los clusters U y V (más próximos, más distantes o en media más próximos) que sea la menor entre todas, d_{uv}
- Juntar U y V en uno solo. Actualizar D:
 - Borrando las filas y columnas de los clusters U y V
 - Formando la fila y columna de las distancias del nuevo cluster (UV) al resto de clusters
- Repetir los pasos (2) y (3) un total de $(N - 1)$ veces

101

101

10. Sist. agrupamiento información jerárquicos aglomerativos

Ejemplo (Univ. Carlos III, J.M. Marin):

- Primera iteración (5 objetos):

$$D = [d_{ik}]_{ik} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{matrix}$$

- La menor distancia (2) hace que se unan 3 y 5
- Distancia entre el cluster (35) y los objetos 1, 2, 4

$$d_{(35),1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$$

$$d_{(35),2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$$

$$d_{(35),4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$$

$$\begin{matrix} (35) & \begin{matrix} 1 & 2 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 4 \end{matrix} & \begin{bmatrix} 0 & & \\ 3 & 0 & \\ 7 & 9 & 0 \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{matrix}$$

102

102

Explotación de la información, Clasificación y Agrupamiento de Información

10. Sist. agrupamiento información jerárquicos aglomerativos

Ejercicio 11:

- Calcular las siguientes iteraciones del algoritmo para calcular el dendograma completo

103

103

Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Tensorflow (Google):

- <https://www.tensorflow.org/>
- TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of **tools, libraries and community resources** that lets researchers push the **state-of-the-art in ML** and developers easily build and deploy ML powered applications.
- **Datasets** collection of ready-to-use datasets:
 - <https://www.tensorflow.org/datasets/catalog/overview>

104

104

Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Tensorflow (cont.):

- Reglas del aprendizaje automático:
 - https://developers.google.com/machine-learning/guides/rules-of-ml/?utm_source=DevSite&utm_campaign=Text-Class-Guide&utm_medium=referral&utm_content=rules-of-ml&utm_term=distribution#training-serving_skew
 - “Regla n.º 1: No tengas miedo de lanzar un producto sin aprendizaje automático”
 - **El aprendizaje automático es genial, pero necesita datos.** En teoría, puedes obtener datos de un problema diferente y, luego, modificar un poco el modelo para un producto nuevo, pero es probable que la heurística básica no funcione como es debido. Si piensas que el aprendizaje automático te brindará un aumento del 100%, entonces una heurística te permitirá alcanzar el 50% de ese camino

105

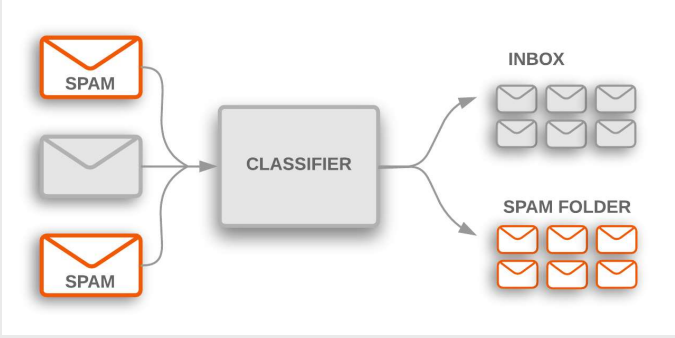
105

Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Tensorflow (cont.):

- Text Classification Workflow
 - <https://developers.google.com/machine-learning/guides/text-classification>



The diagram illustrates a text classification workflow. On the left, three email icons are shown: two labeled 'SPAM' (one orange, one grey) and one unlabeled grey icon. Arrows from these icons point to a central box labeled 'CLASSIFIER'. From the 'CLASSIFIER' box, two arrows point to the right. The top arrow points to a group of six grey envelope icons labeled 'INBOX'. The bottom arrow points to a group of six orange envelope icons labeled 'SPAM FOLDER'.

106

106

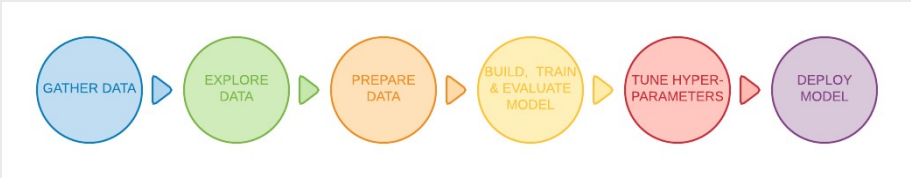
Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Tensorflow (cont.):

■ Text Classification Workflow (cont.):

- "...Gathering data is the most important step in solving any supervised machine learning problem. **Your text classifier can only be as good as the dataset it is built from...**"



107

107

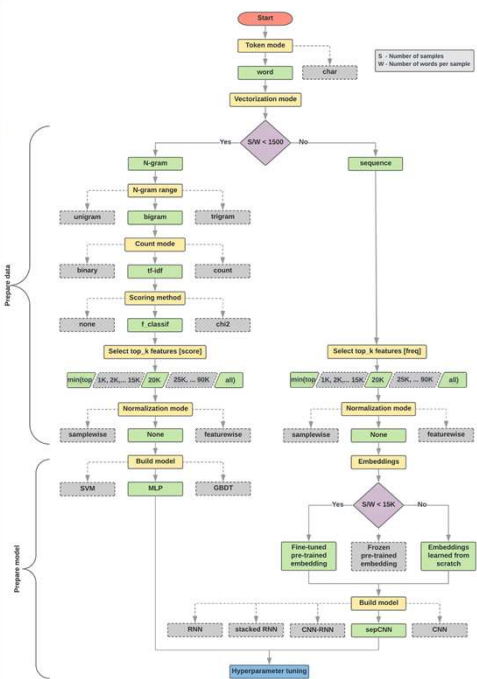
Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Tensorflow (cont.):


■ Text Classification Workflow (cont.)

- Calculate the number of samples/number of words per sample ratio.
- **If this ratio is less than 1500**
 - # Tokenize the text as **n-grams** and use a simple **multi-layer perceptron (MLP)** model to classify them
 - # Else, tokenize the text as sequences and use a **sepCNN** model



108

108



Explotación de la información, Clasificación y Agrupamiento de Información


11. Herramientas

Tensorflow (cont.):

- Text classification with TensorFlow Hub: Movie reviews
 - https://www.tensorflow.org/tutorials/keras/text_classification_with_hub
 - This notebook **classifies movie reviews as positive or negative** using the text of the review
 - We'll use **the IMDB dataset** that contains the text of 50,000 movie reviews from the Internet Movie Database. These are split into 25,000 reviews for training and 25,000 reviews for testing. The **training and testing sets are balanced**, meaning they contain an equal number of positive and negative reviews

109

109



Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Tensorflow (cont.):

- Text Classification Tutorial Pt. 1 (Coding TensorFlow)
 - <https://www.youtube.com/watch?v=BO4g2DRvL6U>
- Text Classification Tutorial Pt. 2 (Coding TensorFlow)
 - <https://www.youtube.com/watch?v=vPrSca-YjFg>

110

110

Explotación de la información, Clasificación y Agrupamiento de Información

11. Herramientas

Información adicional:

- [Ex. Inf. Modulo 4. Deep Learning Based Text Classification_A Comprehensive Review.pdf](#):
 - Jianfeng Gao from Microsoft Research, Nal Kalchbrenner from Google Brain, and Erik Cambria from NTU (2020)
 - <https://arxiv.org/pdf/2004.03705.pdf>

111