

# ACTIVIDAD DE CLASIFICACIÓN DE TEXTO

En esta actividad vamos a trabajar en clasificar textos. Se recorrerá todo el proceso desde traer el dataset hasta proceder a dicha clasificación. Durante la actividad se llevarán a cabo muchos procesos como la creación de un vocabulario, el uso de embeddings y la creación de modelos.

Las cuestiones presentes en esta actividad están basadas en un Notebook creado por François Chollet, uno de los creadores de Keras y autor del libro "Deep Learning with Python".

En este Notebook se trabaja con el dataset "Newsgroup20" que contiene aproximadamente 20000 mensajes que pertenecen a 20 categorías diferentes.

El objetivo es entender los conceptos que se trabajan y ser capaz de hacer pequeñas experimentaciones para mejorar el Notebook creado.

## Librerías

```
import numpy as np
import tensorflow as tf
from tensorflow import keras
```

## Descarga de Datos

```
data_path = keras.utils.get_file(
    "news20.tar.gz",
    "http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.
    tar.gz",
    untar=True,
)

import os
import pathlib

#Estructura de directorios del dataset
data_dir = pathlib.Path(data_path).parent / "20_newsgroup"
dirnames = os.listdir(data_dir)
print("Number of directories:", len(dirnames))
print("Directory names:", dirnames)

Number of directories: 20
Directory names: ['alt.atheism', 'comp.graphics', 'comp.os.ms-
windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware',
'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles',
'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt',
```

```
'sci.electronics', 'sci.med', 'sci.space', 'soc.religion.christian',  
'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc',  
'talk.religion.misc']
```

```
print(data_dir)
```

```
C:\Users\christian.espinosa\.keras\datasets\20_newsgroup
```

```
#Algunos archivos de la categoria "com.graphics"
```

```
fnames = os.listdir(data_dir / "comp.graphics")
```

```
print("Number of files in comp.graphics:", len(fnames))
```

```
print("Some example filenames:", fnames[:5])
```

```
Number of files in comp.graphics: 1000
```

```
Some example filenames: ['37261', '37913', '37914', '37915', '37916']
```

```
#Ejemplo de un texto de la categoría "com.graphics"
```

```
print(open(data_dir / "comp.graphics" / "37261").read())
```

```
Xref: cantaloupe.srv.cs.cmu.edu comp.graphics:37261 alt.graphics:519
```

```
comp.graphics.animation:2614
```

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!ogicse!uwm.edu!
```

```
zaphod.mps.ohio-state.edu!darwin.sura.net!dtix.dt.navy.mil!oasys!
```

```
lipman
```

```
From: lipman@oasys.dt.navy.mil (Robert Lipman)
```

```
Newsgroups: comp.graphics,alt.graphics,comp.graphics.animation
```

```
Subject: CALL FOR PRESENTATIONS: Navy SciViz/VR Seminar
```

```
Message-ID: <32850@oasys.dt.navy.mil>
```

```
Date: 19 Mar 93 20:10:23 GMT
```

```
Article-I.D.: oasys.32850
```

```
Expires: 30 Apr 93 04:00:00 GMT
```

```
Reply-To: lipman@oasys.dt.navy.mil (Robert Lipman)
```

```
Followup-To: comp.graphics
```

```
Distribution: usa
```

```
Organization: Carderock Division, NSWC, Bethesda, MD
```

```
Lines: 65
```

CALL FOR PRESENTATIONS

NAVY SCIENTIFIC VISUALIZATION AND VIRTUAL REALITY SEMINAR

Tuesday, June 22, 1993

Carderock Division, Naval Surface Warfare Center  
(formerly the David Taylor Research Center)  
Bethesda, Maryland

SPONSOR: NESS (Navy Engineering Software System) is sponsoring a one-day Navy Scientific Visualization and Virtual Reality Seminar. The purpose of the seminar is to present and exchange information for

Navy-related scientific visualization and virtual reality programs, research, developments, and applications.

PRESENTATIONS: Presentations are solicited on all aspects of Navy-related scientific visualization and virtual reality. All current work, works-in-progress, and proposed work by Navy organizations will be considered. Four types of presentations are available.

1. Regular presentation: 20-30 minutes in length
2. Short presentation: 10 minutes in length
3. Video presentation: a stand-alone videotape (author need not attend the seminar)
4. Scientific visualization or virtual reality demonstration (BYOH)

Accepted presentations will not be published in any proceedings, however, viewgraphs and other materials will be reproduced for seminar attendees.

ABSTRACTS: Authors should submit a one page abstract and/or videotape to:

Robert Lipman  
Naval Surface Warfare Center, Carderock Division  
Code 2042  
Bethesda, Maryland 20084-5000

VOICE (301) 227-3618; FAX (301) 227-5753  
E-MAIL lipman@oasys.dt.navy.mil

Authors should include the type of presentation, their affiliations, addresses, telephone and FAX numbers, and addresses. Multi-author papers should designate one point of contact.

DEADLINES: The abstract submission deadline is April 30, 1993.  
Notification of acceptance will be sent by May 14, 1993.  
Materials for reproduction must be received by June 1, 1993.

For further information, contact Robert Lipman at the above address.

PLEASE DISTRIBUTE AS WIDELY AS POSSIBLE, THANKS.

Robert Lipman	Internet: lipman@oasys.dt.navy.mil
David Taylor Model Basin - CDNSWC	or: lip@ocean.dt.navy.mil
Computational Signatures and	Voicenet: (301) 227-3618
Structures Group, Code 2042	Factsnet: (301) 227-5753
Bethesda, Maryland 20084-5000	Phishnet: stockings@long.legs

The sixth sick shiek's sixth sheep's sick.

```
import spacy
import en_core_web_sm
nlp = en_core_web_sm.load()
```

```
total_tokens=0
for i in (range(15)):
    doc = nlp(pathlib.Path(data_dir / "comp.graphics" /
fnames[i]).read_text(encoding="latin-1"))
    doc.__len__()
    total_tokens+=doc.__len__()
```

```
average_tokens = total_tokens/15
```

```
print(average_tokens)
```

297.2

```
#Algunos archivos de la categoria "talk.politics.misc"
fnames = os.listdir(data_dir / "talk.politics.misc")
print("Number of files in talk.politics.misc:", len(fnames))
print("Some example filenames:", fnames[:5])
```

```
Number of files in talk.politics.misc: 1000
Some example filenames: ['124146', '176845', '176846', '176847',
'176849']
```

```
#Ejemplo de un texto de la categoría "talk.politics.misc"
print(open(data_dir / "talk.politics.misc" / "178463").read())
```

```
Xref: cantaloupe.srv.cs.cmu.edu talk.politics.guns:54219
talk.politics.misc:178463
Newsgroups: talk.politics.guns,talk.politics.misc
Path: cantaloupe.srv.cs.cmu.edu!magnesium.club.cc.cmu.edu!
news.sei.cmu.edu!cis.ohio-state.edu!magnus.acs.ohio-state.edu!
usenet.ins.cwru.edu!agate!spool.mu.edu!darwin.sura.net!
martha.utcc.utk.edu!FRANKENSTEIN.CE.UTK.EDU!VEAL
From: VEAL@utkvm1.utk.edu (David Veal)
Subject: Re: Proof of the Viability of Gun Control
Message-ID: <VEAL.749.735192116@utkvm1.utk.edu>
Lines: 21
Sender: usenet@martha.utcc.utk.edu (USENET News System)
Organization: University of Tennessee Division of Continuing Education
References: <lqpbqd$ntl@access.digex.net>
<C5otvp.ItL@magpie.linknet.com>
Date: Mon, 19 Apr 1993 04:01:56 GMT
```

[alt.drugs and alt.conspiracy removed from newsgroups line.]

In article <C5otvp.ItL@magpie.linknet.com> neal@magpie.linknet.com  
(Neal) writes:

```
> Once the National Guard has been called into federal service,  
>it is under the command of the present. Tha National Guard, though  
>defined as the "Militia" in the statutes, is actually a reserve  
component  
>of the United State Army, and was formed pursuant to the power of  
Congress  
>to raise and support Armies.
```

That's the really cute thing about saying the 2nd amendment  
only covers the national guard, because that would mean that it  
essentially prohibits the federal government from disarming a branch  
of the federal government.

Sounds like a real limit to federal power to me.

-----  
--

David Veal Univ. of Tenn. Div. of Cont. Education Info. Services Group  
PA146008@utkvm1.utk.edu - "I still remember the way you laughed, the  
day  
your pushed me down the elevator shaft; I'm beginning to think you  
don't  
love me anymore." - "Weird Al"

```
samples = []  
labels = []  
class_names = []  
class_index = 0  
for dirname in sorted(os.listdir(data_dir)):  
    class_names.append(dirname)  
    dirpath = data_dir / dirname  
    fnames = os.listdir(dirpath)  
    print("Processing %s, %d files found" % (dirname, len(fnames)))  
    for fname in fnames:  
        fpath = dirpath / fname  
        f = open(fpath, encoding="latin-1")  
        content = f.read()  
        lines = content.split("\n")  
        lines = lines[10:]  
        content = "\n".join(lines)  
        samples.append(content)  
        labels.append(class_index)  
    class_index += 1  
  
print("Classes:", class_names)  
print("Number of samples:", len(samples))
```

```
Processing alt.atheism, 1000 files found
Processing comp.graphics, 1000 files found
Processing comp.os.ms-windows.misc, 1000 files found
Processing comp.sys.ibm.pc.hardware, 1000 files found
Processing comp.sys.mac.hardware, 1000 files found
Processing comp.windows.x, 1000 files found
Processing misc.forsale, 1000 files found
Processing rec.autos, 1000 files found
Processing rec.motorcycles, 1000 files found
Processing rec.sport.baseball, 1000 files found
Processing rec.sport.hockey, 1000 files found
Processing sci.crypt, 1000 files found
Processing sci.electronics, 1000 files found
Processing sci.med, 1000 files found
Processing sci.space, 1000 files found
Processing soc.religion.christian, 997 files found
Processing talk.politics.guns, 1000 files found
Processing talk.politics.mideast, 1000 files found
Processing talk.politics.misc, 1000 files found
Processing talk.religion.misc, 1000 files found
Classes: ['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc',
'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x',
'misc.forsale', 'rec.autos', 'rec.motorcycles', 'rec.sport.baseball',
'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med',
'sci.space', 'soc.religion.christian', 'talk.politics.guns',
'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc']
Number of samples: 19997
```

## Mezclando los datos para separarlos en Training y Test

```
# Shuffle the data
seed = 1337
rng = np.random.RandomState(seed)
rng.shuffle(samples)
rng = np.random.RandomState(seed)
rng.shuffle(labels)

# Extract a training & validation split
validation_split = 0.2
num_validation_samples = int(validation_split * len(samples))
train_samples = samples[:-num_validation_samples]
val_samples = samples[-num_validation_samples:]
train_labels = labels[:-num_validation_samples]
val_labels = labels[-num_validation_samples:]
```

```

print("Ejemplo de train_samples:")
print(train_samples[0])

print("\nEjemplo de val_samples:")
print(val_samples[0])

print("\nEjemplo de train_labels:")
print(train_labels[0])

print("\nEjemplo de val_labels:")
print(val_labels[0])

print("\nNombre de la clase de train_labels[0]:")
print(class_names[train_labels[0]])

print("\nNombre de la clase de val_labels[0]:")
print(class_names[val_labels[0]])

```

Ejemplo de train\_samples:

In article <1993Apr20.032017.5783@wuecl.wustl.edu> jca2@cec1.wustl.edu (Joseph Charles Achkar) writes:

```

> It was nice to see ESPN show game 1 between the Wings and Leafs
since
>the Cubs and Astros got rained out. Instead of showing another
baseball
>game, they decided on the Stanley Cup Playoffs. A classy move by
ESPN.

```

They tried their best not to show it, believe me. I'm surprised they couldn't find a sprint car race (mini cars through pigpens, indeed!) on short notice.

George

--

George Ferguson	ARPA: ferguson@cs.rochester.edu
Dept. of Computer Science	UUCP: rutgers!rochester!ferguson
University of Rochester	VOX: (716) 275-2527
Rochester NY 14627-0226	FAX: (716) 461-2018

Ejemplo de val\_samples:

Lines: 30

In article <1993Apr22.153528.10877@ra.royalroads.ca>, mlee@post.RoyalRoads.ca (Malcolm Lee) wrote:

```

> Eternal damnation is the consequence of the choice one makes in
rejecting
> God. If you choose to jump off a cliff, you can hardly blame God
for you
> going *splat* at the bottom. He knows that if you choose to jump,

```

that  
> you will die but He will not prevent you from making that choice.  
In fact,  
> He sent His Son to stand on the edge of the cliff and tell everyone  
of what  
> lies below. To prove that point, Jesus took that plunge Himself but  
He being  
> God was able to rise up again. I have seen the example of Christ  
and have  
> chosen not to jump and I'm trying to tell you not to jump or else  
you'll  
> go \*splat\*.  
>  
> You don't have to listen to me and I won't stop you if you decide to  
jump.  
> I only ask that you check it out before taking the plunge. You owe  
it to  
> yourself. I don't like seeing anyone go \*splat\*.

I'm for the moment interested in this notion of the 'leap of faith'  
established by Kierkegaard. It clearly points out a possible solution  
to transcendental values. What I don't understand is that it also  
clearly shows the existentialism system where any leap to any  
transcendental direction is equal.

In other words I might not jump off the cliff mentioned above,  
but at the same time I will decide to what direction I will go.  
Actually I will do it just now.

Cheers,  
Kent

---

sandvik@newton.apple.com. ALink: KSAND -- Private activities on the  
net.

Ejemplo de train\_labels:  
10

Ejemplo de val\_labels:  
19

Nombre de la clase de train\_labels[0]:  
rec.sport.hockey

Nombre de la clase de val\_labels[0]:  
talk.religion.misc

print(train\_samples[:3])



["\nIn article <1993Apr20.032017.5783@wuecl.wustl.edu>  
jca2@cecl.wustl.edu (Joseph Charles Achkar) writes:\n> It was nice to  
see ESPN show game 1 between the Wings and Leafs since\n>the Cubs and  
Astros got rained out. Instead of showing another baseball\n>game,  
they decided on the Stanley Cup Playoffs. A classy move by ESPN.\n\nThey tried their best not to show it, believe me. I'm surprised they\ncouldn't find a sprint car race (mini cars through pigpens, indeed!)\non short notice.\n\nGeorge\n-- \nGeorge Ferguson  
ARPA: ferguson@cs.rochester.edu\nDept. of Computer Science UUCP:  
rutgers!rochester!ferguson\nUniversity of Rochester VOX:  
(716) 275-2527\nRochester NY 14627-0226 FAX: (716) 461-2018\n", "Approved: christian@aramis.rutgers.edu\n\nMy girlfriend is a  
smoker. She has been addicted to it for quite some time.\nShe has been  
tried a couple of times, but then always get back to it. Her \nbackground is non-Christian, but she's interested in Christianity.  
I'm a\nChristian and non-smoker.\n\nI would like to collect any  
personal stories from Christians who managed to \nquit. I hope that  
this will encourage her to keep on trying. If anybody ever \nhad a  
similar problem or knows a good book on it, pls reply by email. \n\nI  
appreciate any kinds of helps. Thanks a lot.\n\n=====

==\nSimon Darjadi Wibowo Telp : (65)7726863\nDept.  
of EE, Nat'l Univ. Of S'pore Fax : (65)7773117\nSingapore 0511  
Internet : eledw@nuscc.nus.sg\n", '\nIn article  
<C5otox.BJI@dscomsa.desy.de> hallam@dscomsa.desy.de (Phill Hallam-  
Baker) writes:\n>\n>In article <VEAL.740.735074621@utkvml1.utk.edu>,  
VEAL@utkvml1.utk.edu (David Veal) writes:\n>\n>|>In article  
<C5n4wH.Izv@dscomsa.desy.de> hallam@dscomsa.desy.de (Phill Hallam-  
Baker) writes:\n>|>>\n>|> Or are they simply propoganda? We  
can't know what Phill \*really\*\n>|>means because he's obviously  
using arguments designed to convince.\n>\n>I make no secret of what I  
am up to, I have stated explicitly in posts\n>that I am a political  
propagandist on numerous occasions. Anyone posting\n>to this group who  
is not probably has the wrong group.\n>\n>For example I have on  
numerous occasions stated quite clearly that I\n>beleive that certain  
factions of the gun lobby are the worst possible\n>advocates of their  
cause and I am prepared to do anything in my power\n>to provide them  
with a platform because they can convince people far\n>better than I  
could hope that many people with a fixation on lethal\n>weapons are  
dangerous and derranged. \n\n If you happen to know a political  
position which does not\n>have people advocating it who do more harm  
than good, please point it\n>nout.\n\n \n>Some people have even  
accused me of inventing such advocates purely \n>for the purpose of  
having them trash a set of political views. In \n>fact this would be a  
futile tactic because I could never hope to\n>invent a character as  
dangerous as sybok.athena.edu, a man who I\n>quite seriously believe  
to be mentaly ill and a potential psychopath.\n>Unfortunately the  
local sherifs office have informed me that they\n>are unable to act  
untill he attacks someone.\n\n One of the advantages and draw-  
backs of requiring proof\nnon the part of the government before they

may take action against citizens. (and part of the reason some of us believe weapons should be available.)

So if you were to ask me what is the point that I am trying to make from this current argument on the absolute sanctity of the US constitution what would I answer?

We are not arguing the absolute sanctity of the U.S. Constitution. In fact, the filibuster we're talking about isn't in the Constitution. I objected to your suggestion that the Senate wasn't intended to exercise the power it was clearly given.

Firstly I see that the current US political scene like the UK political scene has become tied to special interests. Rather than chase the convenient caricatures put about by the media and politicians themselves for this - Gay rights campaigners, environmentalists, zionists (i.e. Jews), "foreign lobbyists" - whatever voting power they have etc. I suggest that you look at who is really benefiting. The inevitable conclusion is that it is the major corporations owned by the ultra-wealthy that have benefited. Reagan and Bush created what can only be described as a welfare state for the rich.

I'll point out again that Reagan only had a Republican Majority in the Senate during his first term, and his coalition in the House came apart at about the same time. Bush never had any real support in Congress.

The real point is that everybody, everywhere got their pork, from the big corporations to the guy I saw last night leaving a convenience store with an armful of junk-food he'd bought with food stamps. (He spent more in food stamps on junk than I make in a week and I'm not on government assistance.)

Money was diverted from programs addressing social needs and poured into the weapons industry in the form of cost-plus profits contracts.

Lessee, let's pull out the old Almanac.

In 1980, total U.S. government budget outlays were 590.9 billion dollars. In 1992 (est) they were 1.4754 trillion dollars, an increase of approx. 884 billion dollars.

In 1980, National Defense cost 133.9 billion dollars. In 1992 it was 307 billion dollars, and increase of 174 billion dollars. That leaves an increase of 710 billion dollars unaccounted for. (This represented an increase of 230%)

In 1980, Income Security (which includes retirement programs, Housing Assistance, and unemployment benefits, and I believe welfare) cost 86.5 billion dollars. In 1992 it was 198 billion dollars, or more than national defense started. (This represented an increase of 230%)

In 1980, the Federal Government spent 32 billion dollars on Medicare. In 1992 they spent 118 billion dollars. (an increase of 368%)

In 1980, the Feds spent 9 billion dollars on housing credits and subsidies of that like. In 1992 it was 87 billion.

In 1980, Health care services and research was 23 billion dollars. In 1992, it was 94 billion dollars.

Agriculture, up 9 billion to 17 billion.

Science, up 11 billion to 16 billion.

Resource conservation up 7 billion to 20 billion.

Education up 14 billion to 45 billion.

Veteran benefits up 12 billion to 33 billion.

Transportation up 13 billion to 34 billion.

About the only things I see which was seriously decreased was

under the Energy category, primarily under "Supply," and "Community Development," in the area of "disaster relief," and between the two of them represent a loss of less than 11 billion dollars.

Where *was* this huge diversion?

In order to rectify this situation there must be constitutional revision. Not that's a stretch. If the current government was pushed by the President to create this mess, wouldn't one expect it to begin to nequalize once the pressure is gone?

Secondly the form of this revision must take account of the changed circumstances of the role of the Federal government. Only assuming that the new role is a positive role we want to continue. I see very little positive about it.

The constitution cannot be used to frustrate the democratic process. The Constitution was *designed* to frustrate the democratic process, so that the voters could be absolutely sure they were getting what they wanted by the time it happened. Nor do I see putting the brakes on the "democratic process" an inherently bad thing. California's riding the edge and every time they pull their ballot initiative nonsense it gets worse.

If the people want to have welfare spending by the federal government they will have it. Sometimes, or perhaps most of the time, the people should be told, "no," and pointed to their local government.

Attempting to prevent this through constitutional trickery only leads to the constitution being brought into disrepute.

Phill, would you do me the very great favor of repeating that in talk.politics.guns?

Methods will always be found to bypass such provisions and once the government gets used to bypassing those provisions they will bypass the others up to the first ammendment.

Cute. We can eliminate violations of the law by eliminating the law.

This is a major reason why the right to own guns should be excluded, the implication that this right is equal to the right to free speech is dangerous.

Free speech alone is dangerous, Phill.

People know that mass ownership of lethal weapons causes thousands of murders a year, the dangerous conclusion they may reach is that the first ammendment may also be the same dangerous mistake.

OK, Phill. All you gotta show me is a clear pattern of *reduction* in homicide rates across several countries and that'll be it. (Not current, mind, you, reduction.)

Note however that this is not the slippery slope argument. It is because the right freedom of speech has been chained to the privilege to own weaponry that the danger arises. The advocates of this privilege must not be allowed to chain freedom of speech to their cause such that if they fall freedom of speech falls as well. Such actions are not the actions of people genuinely interested in freedom.

Who's chaining anything to freedom of speech? By *calling* it a freedom?

Thirdly and most importantly I want to discover a mechanism whereby I can engender intellectual debate as opposed to totemic debate. I consider the grave threat to civilisation to be the loss of the ability to reason about the political debate at anything other than the superficial level. The objection I raise to your basing your case entirely on

the\n>assertion of the supremacy of the US constitution is that the currency\n>of your argument is limited to the currency of the totem upon which it\n>is based. The danger of totems is that they can be reinterpreted in\n>different ways by different people. \n\nPhill, you're a master of subtly changing the subject. I haven't\n\*based\* my argument against raw democracy on the Constitution. I've\ntried to explain why it isn't a good idea. The only time I've referred\ninto the Constitution is to point out it doesn't contain the restrictions\non the veto and the Senate you appear to believe were "meant," but\njust didn't make it in there.\n\nThe Constitution doesn't \*contain\* the 41% filibuster rule.\nI only believe that the rule is a good idea. You can't dismiss that\nas venerating the Constitution because it isn't \*in\* the Constitution.\n-----\n---\nDavid Veal Univ. of Tenn. Div. of Cont. Education Info. Services Group\nPA146008@utkvm1.utk.edu - "I still remember the way you laughed, the day\nyour pushed me down the elevator shaft; I'm beginning to think you don't\nlove me anymore." - "Weird Al"\n']

```
print(val_samples[:1])
```

```
["Lines: 30\n\nIn article <1993Apr22.153528.10877@ra.royalroads.ca>,\nmlee@post.RoyalRoads.ca (Malcolm Lee) wrote:\n> Eternal damnation is the consequence of the choice one makes in rejecting\n> God. If you choose to jump off a cliff, you can hardly blame God for you\n> going *splat* at the bottom. He knows that if you choose to jump, that\n> you will die but He will not prevent you from making that choice. In fact,\n> He sent His Son to stand on the edge of the cliff and tell everyone of what\n> lies below. To prove that point, Jesus took that plunge Himself but He being\n> God was able to rise up again. I have seen the example of Christ and have\n> chosen not to jump and I'm trying to tell you not to jump or else you'll\n> go *splat*.\n> \n> You don't have to listen to me and I won't stop you if you decide to jump.\n> I only ask that you check it out before taking the plunge. You owe it to\n> yourself. I don't like seeing anyone go *splat*.\n\nI'm for the moment interested in this notion of the 'leap of faith'\nestablished by Kierkegaard. It clearly points out a possible solution\nto transcendental values. What I don't understand is that it also\nclearly shows the existentialism system where any leap to any\ntranscendental direction is equal. \n\nIn other words I might not jump off the cliff mentioned above,\nbut at the same time I will decide to what direction I will go.\nActually I will do it just now.\n\nCheers,\nKent\n---\nsandvik@newton.apple.com. ALink: KSAND -- Private activities on the net.\n"]
```

```
print(train_labels[:1])
```

```
[10]
```

```
print(val_labels[:1])
```

[19]

# Tokenización de las palabras con TextVectorization

```
from tensorflow.keras.layers import TextVectorization
vectorizer = TextVectorization(max_tokens=20000,
output_sequence_length=200)
text_ds = tf.data.Dataset.from_tensor_slices(train_samples).batch(128)
vectorizer.adapt(text_ds)

vectorizer.get_vocabulary()[:5]

['', '[UNK]', 'the', 'to', 'of']

len(vectorizer.get_vocabulary())

20000
```

## Viendo la salida de Vectorizer

[illegible]

```

0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
0,
    0,    0]]], dtype=int64)>

voc = vectorizer.get_vocabulary()
word_index = dict(zip(voc, range(len(voc))))

test = ["the", "cat", "sat", "on", "the", "mat"]
[word_index[w] for w in test]

[2, 3457, 1682, 15, 2, 5776]

```

## Tokenización de los datos de entrenamiento y validación

```

x_train = vectorizer(np.array([[s] for s in train_samples])).numpy()
x_val = vectorizer(np.array([[s] for s in val_samples])).numpy()

y_train = np.array(train_labels)
y_val = np.array(val_labels)

```

## Creación y entrenamiento del modelo

```

import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers

def create_model(vocab_size, sequence_length, num_classes):

```

```

    model = keras.Sequential()
    model.add(layers.Embedding(input_dim=vocab_size, output_dim=128,
input_length=sequence_length))
    model.add(layers.GlobalAveragePooling1D())
    model.add(layers.Dense(64, activation='relu'))
    model.add(layers.Dense(num_classes, activation='softmax'))
    return model

```

```

vocab_size = 20000
sequence_length = 200
num_classes = len(set(train_labels))

```

```

model = create_model(vocab_size, sequence_length, num_classes)
model.compile(optimizer='adam',
loss='sparse_categorical_crossentropy', metrics=['accuracy'])

```

```

model.fit(x_train, y_train, epochs=10, validation_data=(x_val, y_val))

```

Epoch 1/10

```

c:\Users\christian.espinosa\Documents\UNIR\TFM\receipt_reader\env\Lib\
site-packages\keras\src\layers\core\embedding.py:90: UserWarning:
Argument `input_length` is deprecated. Just remove it.
    warnings.warn(

```

```

500/500 _____ 9s 15ms/step - accuracy: 0.1273 - loss:
2.8337 - val_accuracy: 0.3828 - val_loss: 2.0247

```

Epoch 2/10

```

500/500 _____ 8s 15ms/step - accuracy: 0.4847 - loss:
1.7605 - val_accuracy: 0.6272 - val_loss: 1.3037

```

Epoch 3/10

```

500/500 _____ 7s 15ms/step - accuracy: 0.6954 - loss:
1.0882 - val_accuracy: 0.7254 - val_loss: 0.9789

```

Epoch 4/10

```

500/500 _____ 7s 14ms/step - accuracy: 0.7813 - loss:
0.7725 - val_accuracy: 0.7697 - val_loss: 0.8075

```

Epoch 5/10

```

500/500 _____ 7s 14ms/step - accuracy: 0.8445 - loss:
0.5704 - val_accuracy: 0.7749 - val_loss: 0.7491

```

Epoch 6/10

```

500/500 _____ 7s 15ms/step - accuracy: 0.8691 - loss:
0.4697 - val_accuracy: 0.7542 - val_loss: 0.7853

```

Epoch 7/10

```

500/500 _____ 8s 15ms/step - accuracy: 0.8818 - loss:
0.4065 - val_accuracy: 0.7664 - val_loss: 0.7306

```

Epoch 8/10

```
500/500 _____ 8s 16ms/step - accuracy: 0.9020 - loss:
0.3322 - val_accuracy: 0.7927 - val_loss: 0.6884
Epoch 9/10
500/500 _____ 8s 15ms/step - accuracy: 0.9106 - loss:
0.2969 - val_accuracy: 0.7874 - val_loss: 0.7004
Epoch 10/10
500/500 _____ 7s 15ms/step - accuracy: 0.9212 - loss:
0.2459 - val_accuracy: 0.7832 - val_loss: 0.7282

<keras.src.callbacks.history.History at 0x2354b5fe290>
```

## Evaluación

```
string_input = keras.Input(shape=(1,), dtype="string")
x = vectorizer(string_input)
preds = model(x)
end_to_end_model = keras.Model(string_input, preds)

probabilities = end_to_end_model.predict(tf.convert_to_tensor([["this
message is about computer graphics and 3D modeling"]],
dtype=tf.string))
class_names[np.argmax(probabilities[0])]

1/1 _____ 0s 248ms/step

'comp.graphics'

probabilities =
end_to_end_model.predict(tf.convert_to_tensor([["politics and federal
courts law that people understand with politician and elects
congressman"]], dtype=tf.string))
class_names[np.argmax(probabilities[0])]

1/1 _____ 0s 32ms/step

'talk.politics.guns'

probabilities = end_to_end_model.predict(tf.convert_to_tensor([["we
are talking about religion"]], dtype=tf.string))
class_names[np.argmax(probabilities[0])]

1/1 _____ 0s 15ms/step

'comp.sys.mac.hardware'
```