Clasificación de reseñas de productos

Proyecto Final Redes Neuronales

Equipo

Molina Bis Victor Hugo

Rivas León Alexis

Sánchez Segura Cristian Alonso

Introducción

Introducción

El comercio por internet consiste en la compra y venta de productos o servicio a través de internet.

Las plataformas cuentan con un sistema de reseñas de productos, en el cual los usuarios pueden escribir comentarios acerca de su experiencia con ellos.

Objetivo

Crear una herramienta que permita clasificar los comentarios de una serie de productos.

La clasificación consiste en dividir los comentarios en dos grupos, los que contienen retroalimentación acerca de los productos y los que no la contienen.

Se pretende ayudar a los vendedores conociendo las recomendaciones que los usuarios tienen acerca de sus productos.

Dataset

```
datos=pd.read_csv('/instrumentos.csv')
datos.head()
```

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A2IBPI20UZIR0U	1384719342	cassandra tu "Yeah, well, that's just like, u...	[0, 0]	Not much to write about here, but it does exac...	5.0	good	1393545600	02 28, 2014
1	A14VAT5EAX3D9S	1384719342	Jake	[13, 14]	The product does exactly as it should and is q...	5.0	Jake	1363392000	03 16, 2013
2	A195EZSQDW3E21	1384719342	Rick Bennette "Rick Bennette"	[1, 1]	The primary job of this device is to block the...	5.0	It Does The Job Well	1377648000	08 28, 2013
3	A2C00NNG1ZQQG2	1384719342	RustyBill "Sunday Rocker"	[0, 0]	Nice windscreen protects my MXL mic and preven...	5.0	GOOD WINDSCREEN FOR THE MONEY	1392336000	02 14, 2014
4	A94QU4C90B1AX	1384719342	SEAN MASLANKA	[0, 0]	This pop filter is great. It looks and perform...	5.0	No more pops when I record my vocals.	1392940800	02 21, 2014

```
[9] datos['resume'] = datos['reviewText']+ '' + datos['summary']
    del datos['reviewText']
    del datos['summary']
```

```
[10] datos.head()
```

	overall	resume
0	0	Not much to write about here, but it does exac...
1	0	The product does exactly as it should and is q...
2	0	The primary job of this device is to block the...
3	0	Nice windscreen protects my MXL mic and preven...
4	0	This pop filter is great. It looks and perform...

Construcción del clasificador

Construcción del clasificador

Se decidió hacer una limpieza del texto como primer paso. Esta limpieza consiste en convertir todo el texto a minúsculas, eliminar números y signos de puntuación.

```
datos["resume"] = datos["resume"].apply(clean_text)
datos.head(10)
```

	overall	resume
0	0	not much to write about here but it does exact...
1	0	the product does exactly as it should and is q...
2	0	the primary job of this device is to block the...
3	0	nice windscreen protects my mxl mic and preven...
4	0	this pop filter is great it looks and performs...
5	0	so good that i bought another one love the he...
6	0	i have used monster cables for years and with ...
7	1	i now use this cable to run from the output of...
8	0	perfect for my epiphone sheraton ii monster c...
9	0	monster makes the best cables and a lifetime w...

Construcción del clasificador

Con **TF** se hace un score de las palabras dentro del documento, mientras más se use una palabra más grande será su **TF**.

IDF es un coeficiente que disminuye a partir del número de ocurrencias de un término en el texto.

Construcción del clasificador

Utilizamos 3 algoritmos de clasificación:

- **Regresión Logística**
- **SVM (Support Vector Machine)**
- **Naive Bayes**

Resultados

Regresión Logística

El desempeño con **Logistic Regression**

```
[131] predicted_LR = text_clf_LR.predict(resume_test["resume"])
      np.mean(predicted_LR == resume_test["overall"])
```

```
0.8904910366328916
```

Naive Bayes

El desempeño del **Naive Bayes**

```
[129] predicted_NB = text_clf.predict(resume_test["resume"])
      np.mean(predicted_NB == resume_test["overall"])
```

```
0.8803585346843336
```

SVM

El desempeño con **Support Vector Machines**

```
[133] predicted_SVM = text_clf_SVM.predict(resume_test["resume"])
      np.mean(predicted_SVM == resume_test["overall"])
```

```
0.8803585346843336
```

Mejores Resultados

SVM

El mejor socre y parámetros para **Support Vector Machine**

```
[ ] print(gs_clf_SVM.best_score_)
    print(gs_clf_SVM.best_params_)
```

```
0.8990253411306043
{'clf-svm__alpha': 1e-05, 'clf-svm__epsilon': 0.01, 'tfidf__use_idf': True, 'vect__ngram_range': (1, 2)}
```

SVM con el conjunto test

```
[ ]  text_clf_SVM = Pipeline([
         ('vect', CountVectorizer(ngram_range=(1,2))),
         ('tfidf', TfidfTransformer()),
         ('clf-svm', SGDClassifier(alpha=1e-05, epsilon=0.01)),
     ])
     text_clf_SVM = text_clf_SVM.fit(resume_train["resume"], resume_train["overall"])
```

Con los nuevos parámetros, se alcanza un 90% de accuracy en el test

```
[ ]  predicted_SVM = text_clf_SVM.predict(resume_test["resume"])
     np.mean(predicted_SVM == resume_test["overall"])
```

```
0.9029618082618862
```

Función clasificatoria

Se crea una función para clasificar reviews

```
def test_sentence(model,sentence):
    sentence = [sentence]
    result = model.predict(sentence)
    res = ['No Feedback','Feedback']
    print("El review es %s" % (res[int(result)]))

test_sentence(text_clf_SVM, "It’s a perfect starter pack. And the price is right.")
test_sentence(text_clf_SVM, "The trumpet is hard to blow and the keys stick")
test_sentence(text_clf_SVM, "This instrument has a good sound and is easy to play. It is good for beginners, especially at its price level.")
test_sentence(text_clf_SVM, "Not worth the money. It isn't the easiest to tune.")
test_sentence(text_clf_SVM, "Lovely recorder. Perfect for my 5 year old who is starting lessons.")
test_sentence(text_clf_SVM, "These instruments (the ones made in Italy, not China) have been great for my preschoolers. We originally bought a trumpet for my son, it held up to s
test_sentence(text_clf_SVM, "I wish I would not have bought this. This sound so horrible, and my kid keeps playing it. I guess you get what you pay for.")
test_sentence(text_clf_SVM, "The piano is very entertaining for kids. It has so many options to play like different instruments . It’s very good. In built music is nice for young
test_sentence(text_clf_SVM, "Piano stopped working even after replacing the batteries few times .The sound quality was ok. Ended up returning")
```

```
El review es No Feedback
El review es Feedback
El review es No Feedback
El review es Feedback
El review es No Feedback
El review es No Feedback
El review es Feedback
El review es No Feedback
El review es Feedback
```

Errores

Se casifican bastante bien los reviews que son buenos, pero se tiene dificultad al clasificar los que contienen feedback

```
[ ] test_sentence(text_clf_SVM, "The recorders arrived on time, but 9 out of 10 had torn/ripped cases. I was not pleased with this as I had bought them for my students, and I had to
    test_sentence(text_clf_SVM, "It's not in perfect tune, unfortunately. It works for kids, but not if you want to play for real.")
    test_sentence(text_clf, "The guitar is much heavier than I thought, so the advantage it’s diminutive size grants for travel, is somewhat diminished by its surprising heft.")
```

```
El review es No Feedback
El review es No Feedback
El review es No Feedback
```

Conclusiones

Conclusiones

Construir un clasificador de texto con un buen desempeño no resulta tan complicado, pero se depende bastante de un buen conjunto de datos.

Los datos no siempre se encuentran como los necesitan, se debe de manejarlos y adaptarlos a las necesidades del problema.

Referencias

Shaikh, J. (2017, July 23). *Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK*. Recuperado de https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a

Villa, J. G. (2019, February 13). *TF IDF: herramientas para mejorar la relevancia de tus contenidos*. Recuperado de https://useo.es/tf-idf-relevancia/\#Que_es_el_TF_IDF

Singh, A. (2019, November 4). *Sentiment Classifier using Tfidf - DataSeries.* Recuperado de https://medium.com/dataseries/sentiment-classifier-using-tfidf-3ffce3f1cbd5