Задание

Решение задачи классификации текстов.

Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета (кроме примера, который рассматривался в лекции). Классификация может быть бинарной или многоклассовой. Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

Необходимо сформировать два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer.

В качестве классификаторов необходимо использовать два классификатора по варианту для Вашей группы.

Вариант

Учебная группа	Классификатор №1	Классификатор №2
ИУ5-22М	RandomForestClassifier	LogisticRegression

Ход выполнения работы

In [14]:

```
from sklearn.datasets import fetch_20newsgroups
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
```

In [2]:

```
категории = ['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
новости_обучение = fetch_20newsgroups(subset='train', categories=категории, shuffle=True, random_state=42)

print("Целевые классы набора данных:", новости_обучение.target_names)
print("\nКоличество образцов:", len(новости_обучение.data))
print("\nНекоторые примеры образцов:\n")
for i in range(3):
    print("Класс:", новости_обучение.target_names[новости_обучение.target[i]])
    print("Текст:", новости_обучение.data[i])
    print("\n-------------------------------------------\n")
```

```
Целевые классы набора данных: ['alt.atheism', 'comp.graphics', 'sci.med', 'soc.religion.christian']

Количество образцов: 2257

Некоторые примеры образцов:

Класс: comp.graphics
Текст: From: sd345@city.ac.uk (Michael Collier)
Subject: Converting images to HP LaserJet III?
Nntp-Posting-Host: hampton
Organization: The City University
Lines: 14

Does anyone know of a good way (standard PC application/PD utility) to
convert tif/img/tga files into LaserJet III format.  We would also like to
do the same, converting to HPGL (HP plotter) files.

Please email any response.

Is this the correct group?

Thanks in advance.  Michael.
-- 
Michael Collier (Programmer)                 The Computer Unit,
Email: M.P.Collier@uk.ac.city                The City University,
Tel: 071 477-8000 x3769                      London,
Fax: 071 477-8565                            EC1V 0HB.


-------------------------------------------
```

Класс: comp.graphics
Текст: From: ani@ms.uky.edu (Aniruddha B. Deglurkar)
Subject: help: Splitting a trimming region along a mesh
Organization: University Of Kentucky, Dept. of Math Sciences
Lines: 28

Hi,

I have a problem, I hope some of the 'gurus' can help me solve.

Background of the problem:
I have a rectangular mesh in the uv domain, i.e the mesh is a
mapping of a 3d Bezier patch into 2d. The area in this domain
which is inside a trimming loop had to be rendered. The trimming
loop is a set of 2d Bezier curve segments.
For the sake of notation: the mesh is made up of cells.

My problem is this :
The trimming area has to be split up into individual smaller
cells bounded by the trimming curve segments. If a cell
is wholly inside the area...then it is output as a whole ,
else it is trivially rejected.

Does any body know how thiss can be done, or is there any algo.
somewhere for doing this.

Any help would be appreciated.

Thanks,
Ani.
--
To get irritated is human, to stay cool, divine.

Класс: soc.religion.christian
Текст: From: djohnson@cs.ucsd.edu (Darin Johnson)
Subject: Re: harrassed at work, could use some prayers
Organization: =CSE Dept., U.C. San Diego
Lines: 63

(Well, I'll email also, but this may apply to other people, so
I'll post also.)

>I've been working at this company for eight years in various
>engineering jobs. I'm female. Yesterday I counted and realized that
>on seven different occasions I've been sexually harrassed at this
>company.

>I dreaded coming back to work today. What if my boss comes in to ask
>me some kind of question...

Your boss should be the person bring these problems to. If he/she
does not seem to take any action, keep going up higher and higher.
Sexual harrassment does not need to be tolerated, and it can be an
enormous emotional support to discuss this with someone and know that
they are trying to do something about it. If you feel you can not
discuss this with your boss, perhaps your company has a personnel
department that can work for you while preserving your privacy. Most
companies will want to deal with this problem because constant anxiety
does seriously affect how effectively employees do their jobs.

It is unclear from your letter if you have done this or not. It is
not inconceivable that management remains ignorant of employee
problems/strife even after eight years (it's a miracle if they do
notice). Perhaps your manager did not bring to the attention of
higher ups? If the company indeed does seem to want to ignore the
entire problem, there may be a state agency willing to fight with
you. (check with a lawyer, a women's resource center, etc to find out)

You may also want to discuss this with your paster, priest, husband,
etc. That is, someone you know will not be judgemental and that is
supportive, comforting, etc. This will bring a lot of healing.

>So I returned at 11:25, only to find that ever single
>person had already left for lunch. They left at 11:15 or so. No one
>could be bothered to call me at the other building, even though my
>number was posted.

```
This happens to a lot of people.  Honest.  I believe it may seem
to be due to gross insensitivity because of the feelings you are
going through.  People in offices tend to be more insensitive while
working than they normally are (maybe it's the hustle or stress or...)
I've had this happen to me a lot, often because they didn't realize
my car was broken, etc.  Then they will come back and wonder why I
didn't want to go (this would tend to make me stop being angry at
being ignored and make me laugh).  Once, we went off without our
boss, who was paying for the lunch :-)

>For this
>reason I hope good Mr. Moderator allows me this latest indulgence.

Well, if you can't turn to the computer for support, what would
we do?  (signs of the computer age :-)

In closing, please don't let the hateful actions of a single person
harm you.  They are doing it because they are still the playground
bully and enjoy seeing the hurt they cause.  And you should not
accept the opinions of an imbecile that you are worthless - much
wiser people hold you in great esteem.
-- 
Darin Johnson
djohnson@ucsd.edu
  - Luxury!  In MY day, we had to make do with 5 bytes of swap...



--------------------------------------------
```

Инициализация CountVectorizer и TfidfVectorizer

In [15]:

```
count_vectorizer = CountVectorizer(stop_words='english')
tfidf_vectorizer = TfidfVectorizer(stop_words='english')
```

Преобразование обучающих данных в признаковые векторы

In [17]:

```
X_count = count_vectorizer.fit_transform(новости_обучение.data)
X_tfidf = tfidf_vectorizer.fit_transform(новости_обучение.data)
```

Вывод формы признаковых векторов

In [19]:

```
print("Форма признакового вектора CountVectorizer:", X_count.shape)
print("Форма признакового вектора TfidfVectorizer:", X_tfidf.shape)
```

```
Форма признакового вектора CountVectorizer: (2257, 35482)
Форма признакового вектора TfidfVectorizer: (2257, 35482)
```

Подготовка меток классов

In [21]:

```
y = новости_обучение.target# Разделение набора данных на обучающий и тестовый
X_count_train, X_count_test, y_train, y_test = train_test_split(X_count, y, test_size=0.2, random_state=42)
X_tfidf_train, X_tfidf_test, y_train, y_test = train_test_split(X_tfidf, y, test_size=0.2, random_state=42)
```

Разделение набора данных на обучающий и тестовый

In [22]:

```
X_count_train, X_count_test, y_train, y_test = train_test_split(X_count, y, test_size=0.2, random_state=42)
X_tfidf_train, X_tfidf_test, y_train, y_test = train_test_split(X_tfidf, y, test_size=0.2, random_state=42)
```

Инициализация классификатора `Random forest`

In [23]:

```
rf_classifier = RandomForestClassifier()
# Обучение классификатора Random forest и предсказание
rf_classifier.fit(X_count_train, y_train)
rf_count_predictions = rf_classifier.predict(X_count_test)
rf_count_report = classification_report(y_test, svc_count_predictions)
```

Инициализация классификатора LogisticRegression

In [25]:

```
lr_classifier = LogisticRegression(max_iter=1000)
```

Обучение классификатора LogisticRegression и предсказание

In [26]:

```
lr_classifier.fit(X_tfidf_train, y_train)
lr_tfidf_predictions = lr_classifier.predict(X_tfidf_test)
lr_tfidf_report = classification_report(y_test, lr_tfidf_predictions)
```

Вывод отчета о классификации

Отчет о классификации классификатора случайный лес

In [27]:

```
svc_count_report
```

Out[27]:

```
'              precision    recall  f1-score   support\n\n           0       0.97      0.87      0.92        86\n           1       0.78      0.98      0.87       107\n           2       0.96      0.83      0.89       132\n
3       0.93      0.94      0.93       127\n\n    accuracy                           0.90       452\n   macro avg       0.91      0.90      0.90       452\nweighted avg       0.91      0.90      0.90       452\n'
```

Отчет о классификации классификатора LogisticRegression

In [28]:

```
lr_tfidf_report
```

Out[28]:

```
'              precision    recall  f1-score   support\n\n           0       0.98      0.92      0.95        86\n           1       0.91      1.00      0.95       107\n           2       0.98      0.95      0.97       132\n
3       0.96      0.94      0.95       127\n\n    accuracy                           0.96       452\n   macro avg       0.96      0.95      0.95       452\nweighted avg       0.96      0.96      0.96       452\n'
```