QINMIN HU

2018年4月3日

LAB 1 预处理报告

一. 实验目的

将数据集正文内容均处理为只有词干的xml文件，且每个xml文件中只有一个文本内容。

二. 实验步骤

1.　解压&提取文档

```
for parent,dirnames,filenames in os.walk(rootdir1_ap):
    for e in filenames:
        filename1.append(e.replace(".gz",""))
    filename1.remove('.DS_Store')
for word in filename1:
    f_name=word
    g_file=gzip.GzipFile(rootdir1_ap+'/'+word+".gz")
    print(g_file)
    open("./disk_AP/"+f_name,"w+").write(g_file.read().decode("utf8","ignore"))
    g_file.close()
    (将.gz文件解压)
def ap_getdata(str_dir,filename):
    dir_filename=str_dir+filename
    ap_open=open(dir_filename,"r")
    file=open(dir_filename+".xml","w")
    ap=""
    for e in ap_open:
        ap=ap+e.replace("&","").replace("#","").replace("@","")
    file.write("<document>\n"+ap+"</document>")
    file.close()
    ap_open.close()
    xmldoc = ET.parse(dir_filename+".xml")
    for e in xmldoc.findall('DOC'):
        for s1 in e.findall('DOCNO'):
            docno_ap.append(s1.text.strip())
        for s7 in e.findall('TEXT'):
            text_ap.append(s7.text)
    ap_processing(str_dir,filename)
```

（先将不标准的xml文件化为只有一个根元素的xml文件——加一个根标签，接着，因为只处理正文部分，所以只提取正文，即以’DOCNO’，’TEXT’作为标签的xml元素）

2. 词条化

```
def tokenization(str_test):
    word_token=[]
    for i in range(len(str_test)):
        if str_test[i]==None:
            word_token.append("None")
        else:
            word_token.append(nltk.word_tokenize(str_test[i]))
    return word_token
```

3. 去除停用词

```
def stopword(str_test):
    test=[]
    for words in str_test:
        s=words[:]
        for word in words:
            if word in stopwords:

                s=[e for e in s if e!=word]
        test.append(s)
    return test
```

4. 词形归并

```
def lemmatization(str_test):
    test=str_test[:]
    lemmatizaer=WordNetLemmatizer()
    for i in range(len(str_test)):
        for j in range(len(str_test[i])):

                if str_test[i][j]=="None":
                    test[i][j]=""
                else:
                    test[i][j]=lemmatizaer.lemmatize(str_test[i][j])
    return test
```

5. 词干还原

```
def stemmed(str_test):
    test=[]
    for words in str_test:
        s=[]
        for word in words:
            p=PorterStemmer()
             （用class PorterStemmer进行词干还原）
        s.append(p.stem(word,0,len(word)-1))
    test.append(s)
return test
```

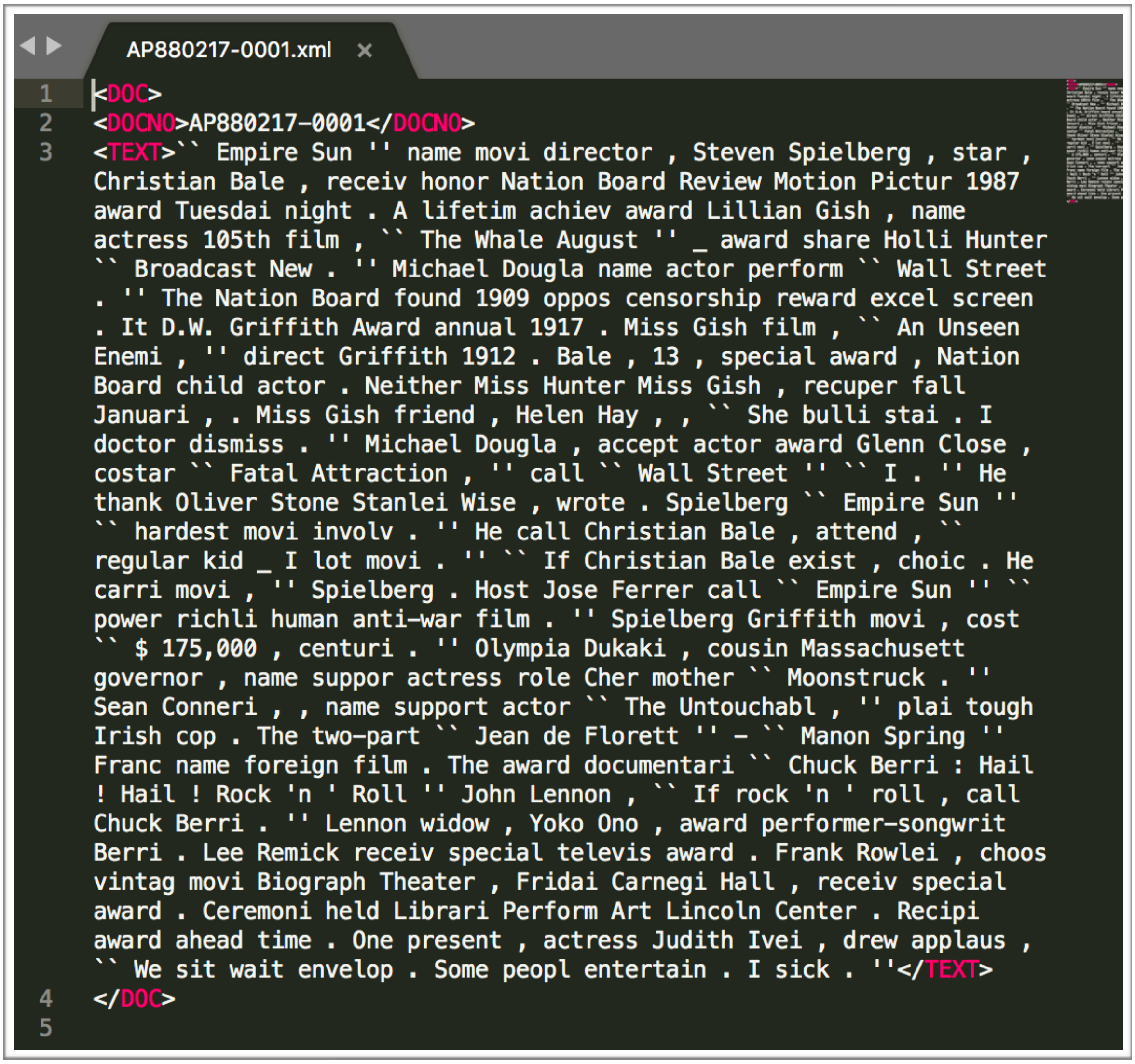
6. 将处理好的文件以xml形式写回&分割

```
def write_file_ap(text,str_dir,filename):
    t=0
    dir_filename=str_dir+filename
    xmldoc = ET.parse(dir_filename+".xml")
    for e in xmldoc.findall('DOC'):
        for s7 in e.findall('TEXT'):
            s7.text=" ".join(text[t])
        for s2 in e.findall('DOCNO'):
            s2.text="".join(docno_ap[t])
        t=t+1
    xmldoc.write(dir_filename+”_new.xml”)
  （写回文件xxx_new.xml）
    flag_text=0
    i=0
    for line in file:
        if "</DOC>" in line:
            test.append(line)
            ap.append(test)
            i=i+1
            test=[]
            flag_text=0
        elif "<DOC>" in line:
            test.append(line)
        elif "<DOCNO>" in line and "</DOCNO>" in line:
            test.append(line)
        elif "<DOCID>" in line and "</DOCID>" in line:
            test.append(line)
        elif "<TEXT>" in line:
            flag_text=1
            test.append(line)
        elif "</TEXT>" in line:
            flag_text=0
            test.append(line)
        elif flag_text==1:
            test.append(line)
        else:
            continue
    s=0
    for words in ap:
        query="".join(docno_ap[s])
        file_test=open('./XML_AP/'+query+".xml",'w')
        word="".join(words)
        file_test.write(word)
        s=s+1
        file_test.close()
    file.close()
```

（分割并写入另一文件夹中）

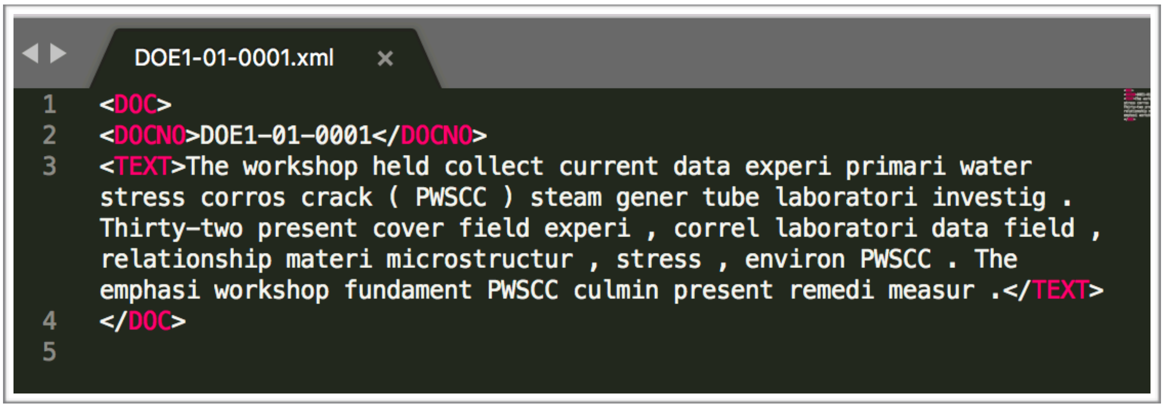
三. 实验结果

AP：



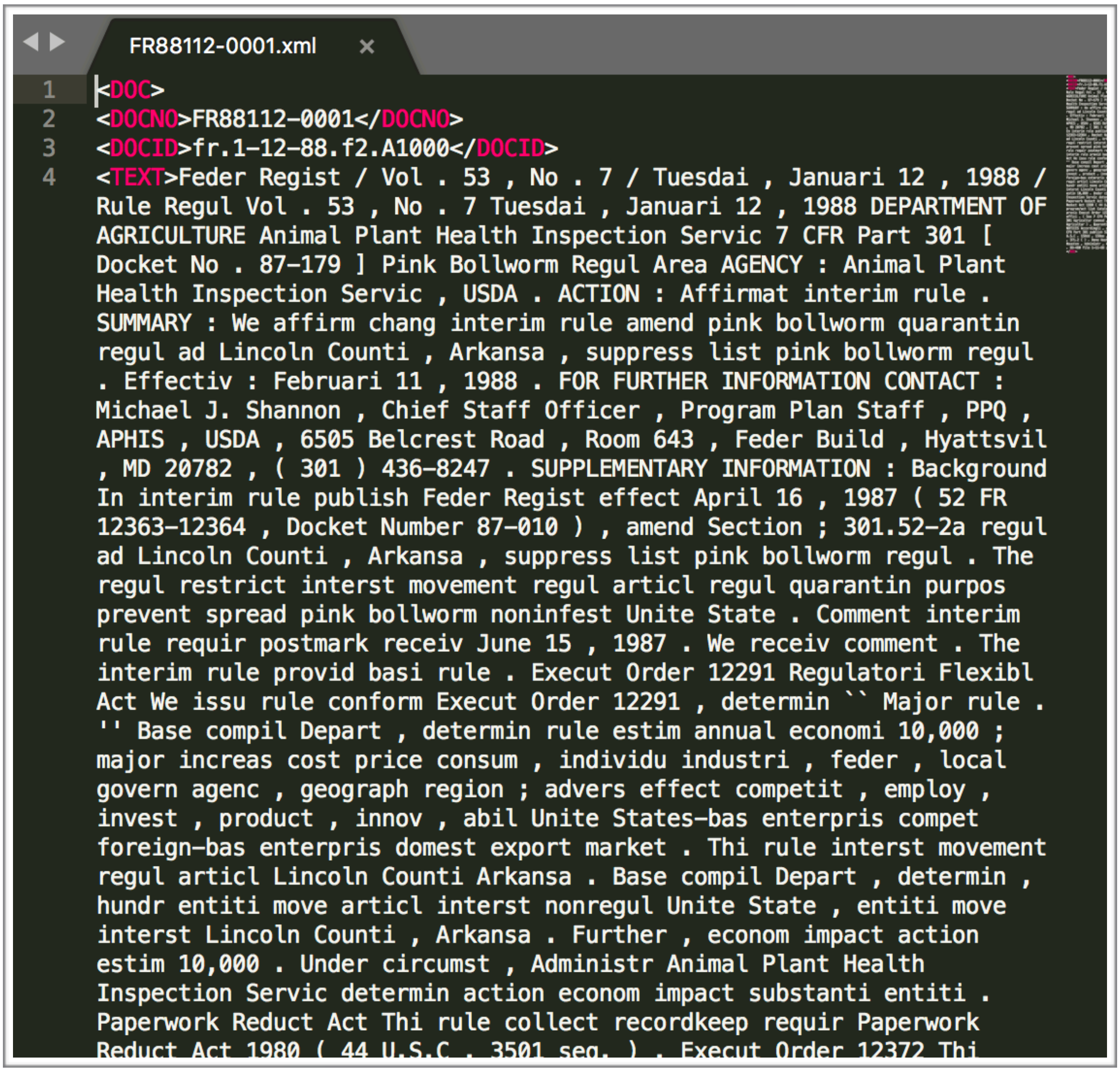
```
<DOC>
<DOCNO>AP880217-0001</DOCNO>
<TEXT>`` Empire Sun '' name movi director , Steven Spielberg , star ,
Christian Bale , receiv honor Nation Board Review Motion Pictur 1987
award Tuesdai night . A lifetim achiev award Lillian Gish , name
actress 105th film , `` The Whale August '' _ award share Holli Hunter
`` Broadcast New . '' Michael Dougla name actor perform `` Wall Street
. '' The Nation Board found 1909 oppos censorship reward excel screen
. It D.W. Griffith Award annual 1917 . Miss Gish film , `` An Unseen
Enemi , '' direct Griffith 1912 . Bale , 13 , special award , Nation
Board child actor . Neither Miss Hunter Miss Gish , recuper fall
Januari , . Miss Gish friend , Helen Hay , , `` She bulli stai . I
doctor dismiss . '' Michael Dougla , accept actor award Glenn Close ,
costar `` Fatal Attraction , '' call `` Wall Street '' `` I . '' He
thank Oliver Stone Stanlei Wise , wrote . Spielberg `` Empire Sun ''
`` hardest movi involv . '' He call Christian Bale , attend , ``
regular kid _ I lot movi . '' `` If Christian Bale exist , choic . He
carri movi , '' Spielberg . Host Jose Ferrer call `` Empire Sun '' ``
power richli human anti-war film . '' Spielberg Griffith movi , cost
`` $ 175,000 , centuri . '' Olympia Dukaki , cousin Massachusett
governor , name suppor actress role Cher mother `` Moonstruck . ''
Sean Conneri , , name support actor `` The Untouchabl , '' plai tough
Irish cop . The two-part `` Jean de Florett '' - `` Manon Spring ''
Franc name foreign film . The award documentari `` Chuck Berri : Hail
! Hail ! Rock 'n ' Roll '' John Lennon , `` If rock 'n ' roll , call
Chuck Berri . '' Lennon widow , Yoko Ono , award performer-songwrit
Berri . Lee Remick receiv special televis award . Frank Rowlei , choos
vintag movi Biograph Theater , Fridai Carnegi Hall , receiv special
award . Ceremoni held Librari Perform Art Lincoln Center . Recipi
award ahead time . One present , actress Judith Ivei , drew applaus ,
`` We sit wait envelop . Some peopl entertain . I sick . ''</TEXT>
</DOC>
```

DOE：



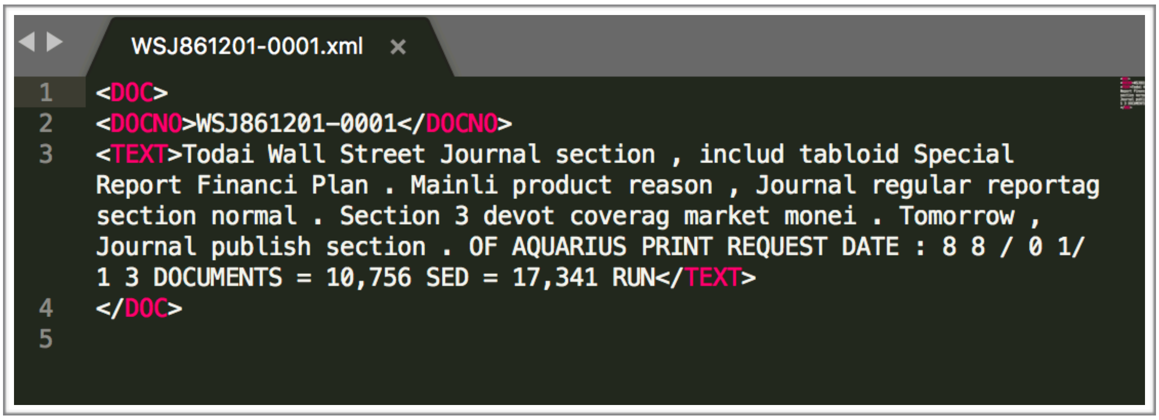
```
<DOC>
<DOCNO>DOE1-01-0001</DOCNO>
<TEXT>The workshop held collect current data experi primari water
stress corros crack ( PWSCC ) steam gener tube laboratori investig .
Thirty-two present cover field experi , correl laboratori data field ,
relationship materi microstructur , stress , environ PWSCC . The
emphasi workshop fundament PWSCC culmin present remedi measur .</TEXT>
</DOC>
```

FR：



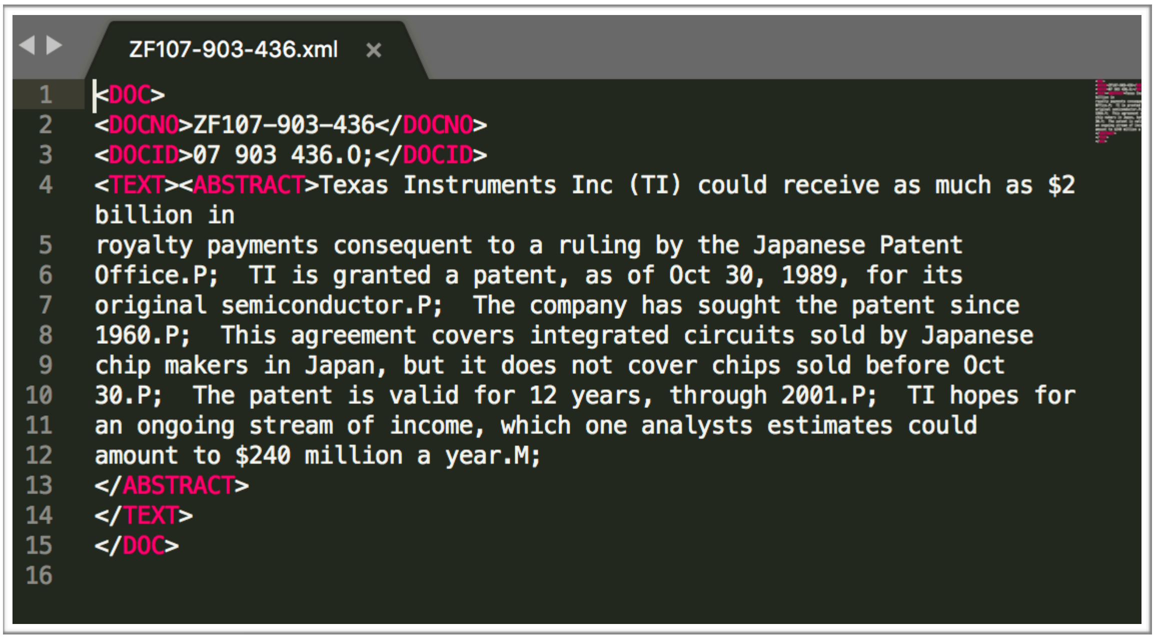
```
<DOC>
<DOCNO>FR88112-0001</DOCNO>
<DOCID>fr.1-12-88.f2.A1000</DOCID>
<TEXT>Feder Regist / Vol . 53 , No . 7 / Tuesdai , Januari 12 , 1988 /
Rule Regul Vol . 53 , No . 7 Tuesdai , Januari 12 , 1988 DEPARTMENT OF
AGRICULTURE Animal Plant Health Inspection Servic 7 CFR Part 301 [
Docket No . 87-179 ] Pink Bollworm Regul Area AGENCY : Animal Plant
Health Inspection Servic , USDA . ACTION : Affirmat interim rule .
SUMMARY : We affirm chang interim rule amend pink bollworm quarantin
regul ad Lincoln Counti , Arkansa , suppress list pink bollworm regul
. Effectiv : Februari 11 , 1988 . FOR FURTHER INFORMATION CONTACT :
Michael J. Shannon , Chief Staff Officer , Program Plan Staff , PPQ ,
APHIS , USDA , 6505 Belcrest Road , Room 643 , Feder Build , Hyattsvil
, MD 20782 , ( 301 ) 436-8247 . SUPPLEMENTARY INFORMATION : Background
In interim rule publish Feder Regist effect April 16 , 1987 ( 52 FR
12363-12364 , Docket Number 87-010 ) , amend Section ; 301.52-2a regul
ad Lincoln Counti , Arkansa , suppress list pink bollworm regul . The
regul restrict interst movement regul articl regul quarantin purpos
prevent spread pink bollworm noninfest Unite State . Comment interim
rule requir postmark receiv June 15 , 1987 . We receiv comment . The
interim rule provid basi rule . Execut Order 12291 Regulatori Flexibl
Act We issu rule conform Execut Order 12291 , determin `` Major rule .
'' Base compil Depart , determin rule estim annual economi 10,000 ;
major increas cost price consum , individu industri , feder , local
govern agenc , geograph region ; advers effect competit , employ ,
invest , product , innov , abil Unite States-bas enterpris compet
foreign-bas enterpris domest export market . Thi rule interst movement
regul articl Lincoln Counti Arkansa . Base compil Depart , determin ,
hundr entiti move articl interst nonregul Unite State , entiti move
interst Lincoln Counti , Arkansa . Further , econom impact action
estim 10,000 . Under circumst , Administr Animal Plant Health
Inspection Servic determin action econom impact substanti entiti .
Paperwork Reduct Act Thi rule collect recordkeep requir Paperwork
Reduct Act 1980 ( 44 U.S.C . 3501 seq. ) . Execut Order 12372 Thi
```

WSJ：



```
<DOC>
<DOCNO>WSJ861201-0001</DOCNO>
<TEXT>Todai Wall Street Journal section , includ tabloid Special
Report Financi Plan . Mainli product reason , Journal regular reportag
section normal . Section 3 devot coverag market monei . Tomorrow ,
Journal publish section . OF AQUARIUS PRINT REQUEST DATE : 8 8 / 0 1/
1 3 DOCUMENTS = 10,756 SED = 17,341 RUN</TEXT>
</DOC>
```

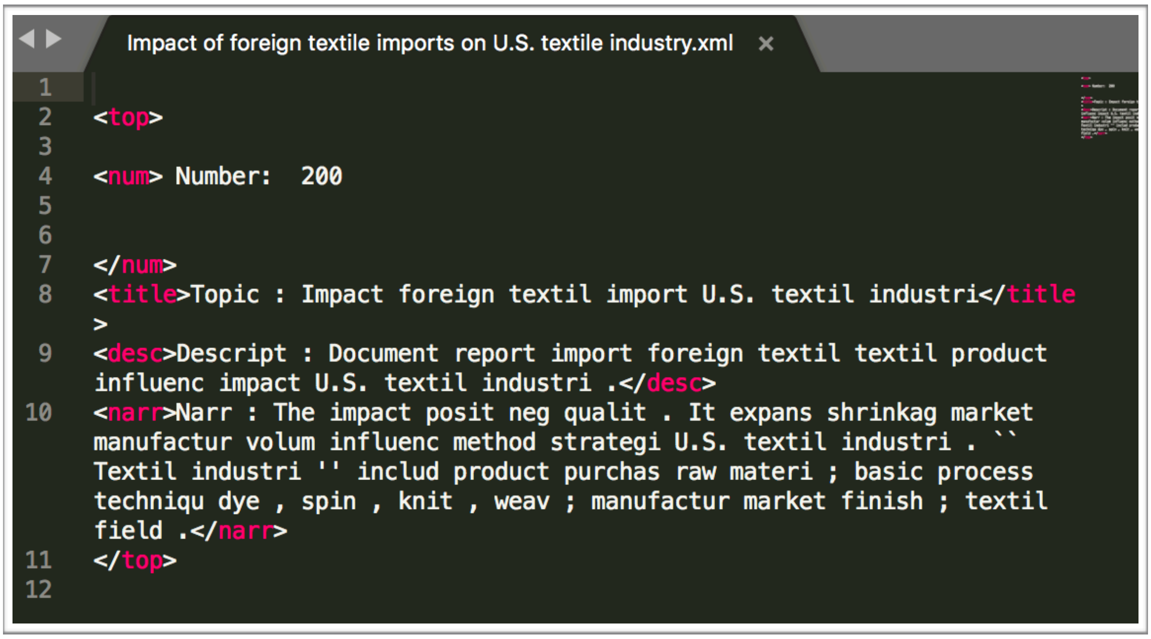
ZF：



```
<DOC>
<DOCNO>ZF107-903-436</DOCNO>
<DOCID>07 903 436.0;</DOCID>
<TEXT><ABSTRACT>Texas Instruments Inc (TI) could receive as much as $2 billion in
royalty payments consequent to a ruling by the Japanese Patent
Office.P;  TI is granted a patent, as of Oct 30, 1989, for its
original semiconductor.P;  The company has sought the patent since
1960.P;  This agreement covers integrated circuits sold by Japanese
chip makers in Japan, but it does not cover chips sold before Oct
30.P;  The patent is valid for 12 years, through 2001.P;  TI hopes for
an ongoing stream of income, which one analysts estimates could
amount to $240 million a year.M;
</ABSTRACT>
</TEXT>
</DOC>

```

topics：



Impact of foreign textile imports on U.S. textile industry.xml

```

<top>

<num> Number:  200


</num>
<title>Topic : Impact foreign textil import U.S. textil industri</title>
<desc>Descript : Document report import foreign textil textil product influenc impact U.S. textil industri .</desc>
<narr>Narr : The impact posit neg qualit . It expans shrinkag market manufactur volum influenc method strategi U.S. textil industri . `` Textil industri '' includ product purchas raw materi ; basic process techniqu dye , spin , knit , weav ; manufactur market finish ; textil field .</narr>
</top>

```