07 - Sentiment-Analysis (movies)

April 24, 2020

1 Sentiment Analysis on movie reviews

Analysis of Social Media Contents

Alessandro Ortis - University of Catania

Based on *Exercise B: Sentiment Analysis* in Natural Language Processing with Python/NLTK by Luciano M. Guasco

1.1 1. Exploring the `movie_reviews` corpus

The nltk.corpus package defines a collection of corpus reader classes, which can be used to access the contents of a diverse set of corpora. The list of available corpora is given at:

http://www.nltk.org/nltk_data/

Each corpus reader class is specialized to handle a specific corpus format. In addition, the nltk.corpus package automatically creates a set of corpus reader instances that can be used to access the corpora in the NLTK data package.

```
In [1]: from nltk.corpus import movie_reviews # These are movie reviews already separated as p
        #movie_reviews.readme().replace('\n', ' ').replace('\t', '').replace('``', '"').replac
```

Most corpora consist of a set of files, each containing a document (or other pieces of text). A list of identifiers for these files is accessed via the fileids() method of the corpus reader.

```
In [2]: movie_reviews.fileids()
```

```
Out[2]: ['neg/cv000_29416.txt',
         'neg/cv001_19502.txt',
         'neg/cv002_17424.txt',
         'neg/cv003_12683.txt',
         'neg/cv004_12641.txt',
         'neg/cv005_29357.txt',
         'neg/cv006_17022.txt',
         'neg/cv007_4992.txt',
         'neg/cv008_29326.txt',
         'neg/cv009_29417.txt',
         'neg/cv010_29063.txt',
         'neg/cv011_13044.txt',
         'neg/cv012_29411.txt',
```

```
'neg/cv013_10494.txt',
'neg/cv014_15600.txt',
'neg/cv015_29356.txt',
'neg/cv016_4348.txt',
'neg/cv017_23487.txt',
'neg/cv018_21672.txt',
'neg/cv019_16117.txt',
'neg/cv020_9234.txt',
'neg/cv021_17313.txt',
'neg/cv022_14227.txt',
'neg/cv023_13847.txt',
'neg/cv024_7033.txt',
'neg/cv025_29825.txt',
'neg/cv026_29229.txt',
'neg/cv027_26270.txt',
'neg/cv028_26964.txt',
'neg/cv029_19943.txt',
'neg/cv030_22893.txt',
'neg/cv031_19540.txt',
'neg/cv032_23718.txt',
'neg/cv033_25680.txt',
'neg/cv034_29446.txt',
'neg/cv035_3343.txt',
'neg/cv036_18385.txt',
'neg/cv037_19798.txt',
'neg/cv038_9781.txt',
'neg/cv039_5963.txt',
'neg/cv040_8829.txt',
'neg/cv041_22364.txt',
'neg/cv042_11927.txt',
'neg/cv043_16808.txt',
'neg/cv044_18429.txt',
'neg/cv045_25077.txt',
'neg/cv046_10613.txt',
'neg/cv047_18725.txt',
'neg/cv048_18380.txt',
'neg/cv049_21917.txt',
'neg/cv050_12128.txt',
'neg/cv051_10751.txt',
'neg/cv052_29318.txt',
'neg/cv053_23117.txt',
'neg/cv054_4101.txt',
'neg/cv055_8926.txt',
'neg/cv056_14663.txt',
'neg/cv057_7962.txt',
'neg/cv058_8469.txt',
'neg/cv059_28723.txt',
'neg/cv060_11754.txt',
```

```
'neg/cv061_9321.txt',
'neg/cv062_24556.txt',
'neg/cv063_28852.txt',
'neg/cv064_25842.txt',
'neg/cv065_16909.txt',
'neg/cv066_11668.txt',
'neg/cv067_21192.txt',
'neg/cv068_14810.txt',
'neg/cv069_11613.txt',
'neg/cv070_13249.txt',
'neg/cv071_12969.txt',
'neg/cv072_5928.txt',
'neg/cv073_23039.txt',
'neg/cv074_7188.txt',
'neg/cv075_6250.txt',
'neg/cv076_26009.txt',
'neg/cv077_23172.txt',
'neg/cv078_16506.txt',
'neg/cv079_12766.txt',
'neg/cv080_14899.txt',
'neg/cv081_18241.txt',
'neg/cv082_11979.txt',
'neg/cv083_25491.txt',
'neg/cv084_15183.txt',
'neg/cv085_15286.txt',
'neg/cv086_19488.txt',
'neg/cv087_2145.txt',
'neg/cv088_25274.txt',
'neg/cv089_12222.txt',
'neg/cv090_0049.txt',
'neg/cv091_7899.txt',
'neg/cv092_27987.txt',
'neg/cv093_15606.txt',
'neg/cv094_27868.txt',
'neg/cv095_28730.txt',
'neg/cv096_12262.txt',
'neg/cv097_26081.txt',
'neg/cv098_17021.txt',
'neg/cv099_11189.txt',
'neg/cv100_12406.txt',
'neg/cv101_10537.txt',
'neg/cv102_8306.txt',
'neg/cv103_11943.txt',
'neg/cv104_19176.txt',
'neg/cv105_19135.txt',
'neg/cv106_18379.txt',
'neg/cv107_25639.txt',
'neg/cv108_17064.txt',
```

```
'neg/cv109_22599.txt',
'neg/cv110_27832.txt',
'neg/cv111_12253.txt',
'neg/cv112_12178.txt',
'neg/cv113_24354.txt',
'neg/cv114_19501.txt',
'neg/cv115_26443.txt',
'neg/cv116_28734.txt',
'neg/cv117_25625.txt',
'neg/cv118_28837.txt',
'neg/cv119_9909.txt',
'neg/cv120_3793.txt',
'neg/cv121_18621.txt',
'neg/cv122_7891.txt',
'neg/cv123_12165.txt',
'neg/cv124_3903.txt',
'neg/cv125_9636.txt',
'neg/cv126_28821.txt',
'neg/cv127_16451.txt',
'neg/cv128_29444.txt',
'neg/cv129_18373.txt',
'neg/cv130_18521.txt',
'neg/cv131_11568.txt',
'neg/cv132_5423.txt',
'neg/cv133_18065.txt',
'neg/cv134_23300.txt',
'neg/cv135_12506.txt',
'neg/cv136_12384.txt',
'neg/cv137_17020.txt',
'neg/cv138_13903.txt',
'neg/cv139_14236.txt',
'neg/cv140_7963.txt',
'neg/cv141_17179.txt',
'neg/cv142_23657.txt',
'neg/cv143_21158.txt',
'neg/cv144_5010.txt',
'neg/cv145_12239.txt',
'neg/cv146_19587.txt',
'neg/cv147_22625.txt',
'neg/cv148_18084.txt',
'neg/cv149_17084.txt',
'neg/cv150_14279.txt',
'neg/cv151_17231.txt',
'neg/cv152_9052.txt',
'neg/cv153_11607.txt',
'neg/cv154_9562.txt',
'neg/cv155_7845.txt',
'neg/cv156_11119.txt',
```

```
'neg/cv157_29302.txt',
'neg/cv158_10914.txt',
'neg/cv159_29374.txt',
'neg/cv160_10848.txt',
'neg/cv161_12224.txt',
'neg/cv162_10977.txt',
'neg/cv163_10110.txt',
'neg/cv164_23451.txt',
'neg/cv165_2389.txt',
'neg/cv166_11959.txt',
'neg/cv167_18094.txt',
'neg/cv168_7435.txt',
'neg/cv169_24973.txt',
'neg/cv170_29808.txt',
'neg/cv171_15164.txt',
'neg/cv172_12037.txt',
'neg/cv173_4295.txt',
'neg/cv174_9735.txt',
'neg/cv175_7375.txt',
'neg/cv176_14196.txt',
'neg/cv177_10904.txt',
'neg/cv178_14380.txt',
'neg/cv179_9533.txt',
'neg/cv180_17823.txt',
'neg/cv181_16083.txt',
'neg/cv182_7791.txt',
'neg/cv183_19826.txt',
'neg/cv184_26935.txt',
'neg/cv185_28372.txt',
'neg/cv186_2396.txt',
'neg/cv187_14112.txt',
'neg/cv188_20687.txt',
'neg/cv189_24248.txt',
'neg/cv190_27176.txt',
'neg/cv191_29539.txt',
'neg/cv192_16079.txt',
'neg/cv193_5393.txt',
'neg/cv194_12855.txt',
'neg/cv195_16146.txt',
'neg/cv196_28898.txt',
'neg/cv197_29271.txt',
'neg/cv198_19313.txt',
'neg/cv199_9721.txt',
'neg/cv200_29006.txt',
'neg/cv201_7421.txt',
'neg/cv202_11382.txt',
'neg/cv203_19052.txt',
'neg/cv204_8930.txt',
```

```
'neg/cv205_9676.txt',
'neg/cv206_15893.txt',
'neg/cv207_29141.txt',
'neg/cv208_9475.txt',
'neg/cv209_28973.txt',
'neg/cv210_9557.txt',
'neg/cv211_9955.txt',
'neg/cv212_10054.txt',
'neg/cv213_20300.txt',
'neg/cv214_13285.txt',
'neg/cv215_23246.txt',
'neg/cv216_20165.txt',
'neg/cv217_28707.txt',
'neg/cv218_25651.txt',
'neg/cv219_19874.txt',
'neg/cv220_28906.txt',
'neg/cv221_27081.txt',
'neg/cv222_18720.txt',
'neg/cv223_28923.txt',
'neg/cv224_18875.txt',
'neg/cv225_29083.txt',
'neg/cv226_26692.txt',
'neg/cv227_25406.txt',
'neg/cv228_5644.txt',
'neg/cv229_15200.txt',
'neg/cv230_7913.txt',
'neg/cv231_11028.txt',
'neg/cv232_16768.txt',
'neg/cv233_17614.txt',
'neg/cv234_22123.txt',
'neg/cv235_10704.txt',
'neg/cv236_12427.txt',
'neg/cv237_20635.txt',
'neg/cv238_14285.txt',
'neg/cv239_29828.txt',
'neg/cv240_15948.txt',
'neg/cv241_24602.txt',
'neg/cv242_11354.txt',
'neg/cv243_22164.txt',
'neg/cv244_22935.txt',
'neg/cv245_8938.txt',
'neg/cv246_28668.txt',
'neg/cv247_14668.txt',
'neg/cv248_15672.txt',
'neg/cv249_12674.txt',
'neg/cv250_26462.txt',
'neg/cv251_23901.txt',
'neg/cv252_24974.txt',
```

```
'neg/cv253_10190.txt',
'neg/cv254_5870.txt',
'neg/cv255_15267.txt',
'neg/cv256_16529.txt',
'neg/cv257_11856.txt',
'neg/cv258_5627.txt',
'neg/cv259_11827.txt',
'neg/cv260_15652.txt',
'neg/cv261_11855.txt',
'neg/cv262_13812.txt',
'neg/cv263_20693.txt',
'neg/cv264_14108.txt',
'neg/cv265_11625.txt',
'neg/cv266_26644.txt',
'neg/cv267_16618.txt',
'neg/cv268_20288.txt',
'neg/cv269_23018.txt',
'neg/cv270_5873.txt',
'neg/cv271_15364.txt',
'neg/cv272_20313.txt',
'neg/cv273_28961.txt',
'neg/cv274_26379.txt',
'neg/cv275_28725.txt',
'neg/cv276_17126.txt',
'neg/cv277_20467.txt',
'neg/cv278_14533.txt',
'neg/cv279_19452.txt',
'neg/cv280_8651.txt',
'neg/cv281_24711.txt',
'neg/cv282_6833.txt',
'neg/cv283_11963.txt',
'neg/cv284_20530.txt',
'neg/cv285_18186.txt',
'neg/cv286_26156.txt',
'neg/cv287_17410.txt',
'neg/cv288_20212.txt',
'neg/cv289_6239.txt',
'neg/cv290_11981.txt',
'neg/cv291_26844.txt',
'neg/cv292_7804.txt',
'neg/cv293_29731.txt',
'neg/cv294_12695.txt',
'neg/cv295_17060.txt',
'neg/cv296_13146.txt',
'neg/cv297_10104.txt',
'neg/cv298_24487.txt',
'neg/cv299_17950.txt',
'neg/cv300_23302.txt',
```

```
'neg/cv301_13010.txt',
'neg/cv302_26481.txt',
'neg/cv303_27366.txt',
'neg/cv304_28489.txt',
'neg/cv305_9937.txt',
'neg/cv306_10859.txt',
'neg/cv307_26382.txt',
'neg/cv308_5079.txt',
'neg/cv309_23737.txt',
'neg/cv310_14568.txt',
'neg/cv311_17708.txt',
'neg/cv312_29308.txt',
'neg/cv313_19337.txt',
'neg/cv314_16095.txt',
'neg/cv315_12638.txt',
'neg/cv316_5972.txt',
'neg/cv317_25111.txt',
'neg/cv318_11146.txt',
'neg/cv319_16459.txt',
'neg/cv320_9693.txt',
'neg/cv321_14191.txt',
'neg/cv322_21820.txt',
'neg/cv323_29633.txt',
'neg/cv324_7502.txt',
'neg/cv325_18330.txt',
'neg/cv326_14777.txt',
'neg/cv327_21743.txt',
'neg/cv328_10908.txt',
'neg/cv329_29293.txt',
'neg/cv330_29675.txt',
'neg/cv331_8656.txt',
'neg/cv332_17997.txt',
'neg/cv333_9443.txt',
'neg/cv334_0074.txt',
'neg/cv335_16299.txt',
'neg/cv336_10363.txt',
'neg/cv337_29061.txt',
'neg/cv338_9183.txt',
'neg/cv339_22452.txt',
'neg/cv340_14776.txt',
'neg/cv341_25667.txt',
'neg/cv342_20917.txt',
'neg/cv343_10906.txt',
'neg/cv344_5376.txt',
'neg/cv345_9966.txt',
'neg/cv346_19198.txt',
'neg/cv347_14722.txt',
'neg/cv348_19207.txt',
```

```
'neg/cv349_15032.txt',
'neg/cv350_22139.txt',
'neg/cv351_17029.txt',
'neg/cv352_5414.txt',
'neg/cv353_19197.txt',
'neg/cv354_8573.txt',
'neg/cv355_18174.txt',
'neg/cv356_26170.txt',
'neg/cv357_14710.txt',
'neg/cv358_11557.txt',
'neg/cv359_6751.txt',
'neg/cv360_8927.txt',
'neg/cv361_28738.txt',
'neg/cv362_16985.txt',
'neg/cv363_29273.txt',
'neg/cv364_14254.txt',
'neg/cv365_12442.txt',
'neg/cv366_10709.txt',
'neg/cv367_24065.txt',
'neg/cv368_11090.txt',
'neg/cv369_14245.txt',
'neg/cv370_5338.txt',
'neg/cv371_8197.txt',
'neg/cv372_6654.txt',
'neg/cv373_21872.txt',
'neg/cv374_26455.txt',
'neg/cv375_9932.txt',
'neg/cv376_20883.txt',
'neg/cv377_8440.txt',
'neg/cv378_21982.txt',
'neg/cv379_23167.txt',
'neg/cv380_8164.txt',
'neg/cv381_21673.txt',
'neg/cv382_8393.txt',
'neg/cv383_14662.txt',
'neg/cv384_18536.txt',
'neg/cv385_29621.txt',
'neg/cv386_10229.txt',
'neg/cv387_12391.txt',
'neg/cv388_12810.txt',
'neg/cv389_9611.txt',
'neg/cv390_12187.txt',
'neg/cv391_11615.txt',
'neg/cv392_12238.txt',
'neg/cv393_29234.txt',
'neg/cv394_5311.txt',
'neg/cv395_11761.txt',
'neg/cv396_19127.txt',
```

```
'neg/cv397_28890.txt',
'neg/cv398_17047.txt',
'neg/cv399_28593.txt',
'neg/cv400_20631.txt',
'neg/cv401_13758.txt',
'neg/cv402_16097.txt',
'neg/cv403_6721.txt',
'neg/cv404_21805.txt',
'neg/cv405_21868.txt',
'neg/cv406_22199.txt',
'neg/cv407_23928.txt',
'neg/cv408_5367.txt',
'neg/cv409_29625.txt',
'neg/cv410_25624.txt',
'neg/cv411_16799.txt',
'neg/cv412_25254.txt',
'neg/cv413_7893.txt',
'neg/cv414_11161.txt',
'neg/cv415_23674.txt',
'neg/cv416_12048.txt',
'neg/cv417_14653.txt',
'neg/cv418_16562.txt',
'neg/cv419_14799.txt',
'neg/cv420_28631.txt',
'neg/cv421_9752.txt',
'neg/cv422_9632.txt',
'neg/cv423_12089.txt',
'neg/cv424_9268.txt',
'neg/cv425_8603.txt',
'neg/cv426_10976.txt',
'neg/cv427_11693.txt',
'neg/cv428_12202.txt',
'neg/cv429_7937.txt',
'neg/cv430_18662.txt',
'neg/cv431_7538.txt',
'neg/cv432_15873.txt',
'neg/cv433_10443.txt',
'neg/cv434_5641.txt',
'neg/cv435_24355.txt',
'neg/cv436_20564.txt',
'neg/cv437_24070.txt',
'neg/cv438_8500.txt',
'neg/cv439_17633.txt',
'neg/cv440_16891.txt',
'neg/cv441_15276.txt',
'neg/cv442_15499.txt',
'neg/cv443_22367.txt',
'neg/cv444_9975.txt',
```

```
'neg/cv445_26683.txt',
'neg/cv446_12209.txt',
'neg/cv447_27334.txt',
'neg/cv448_16409.txt',
'neg/cv449_9126.txt',
'neg/cv450_8319.txt',
'neg/cv451_11502.txt',
'neg/cv452_5179.txt',
'neg/cv453_10911.txt',
'neg/cv454_21961.txt',
'neg/cv455_28866.txt',
'neg/cv456_20370.txt',
'neg/cv457_19546.txt',
'neg/cv458_9000.txt',
'neg/cv459_21834.txt',
'neg/cv460_11723.txt',
'neg/cv461_21124.txt',
'neg/cv462_20788.txt',
'neg/cv463_10846.txt',
'neg/cv464_17076.txt',
'neg/cv465_23401.txt',
'neg/cv466_20092.txt',
'neg/cv467_26610.txt',
'neg/cv468_16844.txt',
'neg/cv469_21998.txt',
'neg/cv470_17444.txt',
'neg/cv471_18405.txt',
'neg/cv472_29140.txt',
'neg/cv473_7869.txt',
'neg/cv474_10682.txt',
'neg/cv475_22978.txt',
'neg/cv476_18402.txt',
'neg/cv477_23530.txt',
'neg/cv478_15921.txt',
'neg/cv479_5450.txt',
'neg/cv480_21195.txt',
'neg/cv481_7930.txt',
'neg/cv482_11233.txt',
'neg/cv483_18103.txt',
'neg/cv484_26169.txt',
'neg/cv485_26879.txt',
'neg/cv486_9788.txt',
'neg/cv487_11058.txt',
'neg/cv488_21453.txt',
'neg/cv489_19046.txt',
'neg/cv490_18986.txt',
'neg/cv491_12992.txt',
'neg/cv492_19370.txt',
```

```
'neg/cv493_14135.txt',
'neg/cv494_18689.txt',
'neg/cv495_16121.txt',
'neg/cv496_11185.txt',
'neg/cv497_27086.txt',
'neg/cv498_9288.txt',
'neg/cv499_11407.txt',
'neg/cv500_10722.txt',
'neg/cv501_12675.txt',
'neg/cv502_10970.txt',
'neg/cv503_11196.txt',
'neg/cv504_29120.txt',
'neg/cv505_12926.txt',
'neg/cv506_17521.txt',
'neg/cv507_9509.txt',
'neg/cv508_17742.txt',
'neg/cv509_17354.txt',
'neg/cv510_24758.txt',
'neg/cv511_10360.txt',
'neg/cv512_17618.txt',
'neg/cv513_7236.txt',
'neg/cv514_12173.txt',
'neg/cv515_18484.txt',
'neg/cv516_12117.txt',
'neg/cv517_20616.txt',
'neg/cv518_14798.txt',
'neg/cv519_16239.txt',
'neg/cv520_13297.txt',
'neg/cv521_1730.txt',
'neg/cv522_5418.txt',
'neg/cv523_18285.txt',
'neg/cv524_24885.txt',
'neg/cv525_17930.txt',
'neg/cv526_12868.txt',
'neg/cv527_10338.txt',
'neg/cv528_11669.txt',
'neg/cv529_10972.txt',
'neg/cv530_17949.txt',
'neg/cv531_26838.txt',
'neg/cv532_6495.txt',
'neg/cv533_9843.txt',
'neg/cv534_15683.txt',
'neg/cv535_21183.txt',
'neg/cv536_27221.txt',
'neg/cv537_13516.txt',
'neg/cv538_28485.txt',
'neg/cv539_21865.txt',
'neg/cv540_3092.txt',
```

```
'neg/cv541_28683.txt',
'neg/cv542_20359.txt',
'neg/cv543_5107.txt',
'neg/cv544_5301.txt',
'neg/cv545_12848.txt',
'neg/cv546_12723.txt',
'neg/cv547_18043.txt',
'neg/cv548_18944.txt',
'neg/cv549_22771.txt',
'neg/cv550_23226.txt',
'neg/cv551_11214.txt',
'neg/cv552_0150.txt',
'neg/cv553_26965.txt',
'neg/cv554_14678.txt',
'neg/cv555_25047.txt',
'neg/cv556_16563.txt',
'neg/cv557_12237.txt',
'neg/cv558_29376.txt',
'neg/cv559_0057.txt',
'neg/cv560_18608.txt',
'neg/cv561_9484.txt',
'neg/cv562_10847.txt',
'neg/cv563_18610.txt',
'neg/cv564_12011.txt',
'neg/cv565_29403.txt',
'neg/cv566_8967.txt',
'neg/cv567_29420.txt',
'neg/cv568_17065.txt',
'neg/cv569_26750.txt',
'neg/cv570_28960.txt',
'neg/cv571_29292.txt',
'neg/cv572_20053.txt',
'neg/cv573_29384.txt',
'neg/cv574_23191.txt',
'neg/cv575_22598.txt',
'neg/cv576_15688.txt',
'neg/cv577_28220.txt',
'neg/cv578_16825.txt',
'neg/cv579_12542.txt',
'neg/cv580_15681.txt',
'neg/cv581_20790.txt',
'neg/cv582_6678.txt',
'neg/cv583_29465.txt',
'neg/cv584_29549.txt',
'neg/cv585_23576.txt',
'neg/cv586_8048.txt',
'neg/cv587_20532.txt',
'neg/cv588_14467.txt',
```

```
'neg/cv589_12853.txt',
'neg/cv590_20712.txt',
'neg/cv591_24887.txt',
'neg/cv592_23391.txt',
'neg/cv593_11931.txt',
'neg/cv594_11945.txt',
'neg/cv595_26420.txt',
'neg/cv596_4367.txt',
'neg/cv597_26744.txt',
'neg/cv598_18184.txt',
'neg/cv599_22197.txt',
'neg/cv600_25043.txt',
'neg/cv601_24759.txt',
'neg/cv602_8830.txt',
'neg/cv603_18885.txt',
'neg/cv604_23339.txt',
'neg/cv605_12730.txt',
'neg/cv606_17672.txt',
'neg/cv607_8235.txt',
'neg/cv608_24647.txt',
'neg/cv609_25038.txt',
'neg/cv610_24153.txt',
'neg/cv611_2253.txt',
'neg/cv612_5396.txt',
'neg/cv613_23104.txt',
'neg/cv614_11320.txt',
'neg/cv615_15734.txt',
'neg/cv616_29187.txt',
'neg/cv617_9561.txt',
'neg/cv618_9469.txt',
'neg/cv619_13677.txt',
'neg/cv620_2556.txt',
'neg/cv621_15984.txt',
'neg/cv622_8583.txt',
'neg/cv623_16988.txt',
'neg/cv624_11601.txt',
'neg/cv625_13518.txt',
'neg/cv626_7907.txt',
'neg/cv627_12603.txt',
'neg/cv628_20758.txt',
'neg/cv629_16604.txt',
'neg/cv630_10152.txt',
'neg/cv631_4782.txt',
'neg/cv632_9704.txt',
'neg/cv633_29730.txt',
'neg/cv634_11989.txt',
'neg/cv635_0984.txt',
'neg/cv636_16954.txt',
```

```
'neg/cv637_13682.txt',
'neg/cv638_29394.txt',
'neg/cv639_10797.txt',
'neg/cv640_5380.txt',
'neg/cv641_13412.txt',
'neg/cv642_29788.txt',
'neg/cv643_29282.txt',
'neg/cv644_18551.txt',
'neg/cv645_17078.txt',
'neg/cv646_16817.txt',
'neg/cv647_15275.txt',
'neg/cv648_17277.txt',
'neg/cv649_13947.txt',
'neg/cv650_15974.txt',
'neg/cv651_11120.txt',
'neg/cv652_15653.txt',
'neg/cv653_2107.txt',
'neg/cv654_19345.txt',
'neg/cv655_12055.txt',
'neg/cv656_25395.txt',
'neg/cv657_25835.txt',
'neg/cv658_11186.txt',
'neg/cv659_21483.txt',
'neg/cv660_23140.txt',
'neg/cv661_25780.txt',
'neg/cv662_14791.txt',
'neg/cv663_14484.txt',
'neg/cv664_4264.txt',
'neg/cv665_29386.txt',
'neg/cv666_20301.txt',
'neg/cv667_19672.txt',
'neg/cv668_18848.txt',
'neg/cv669_24318.txt',
'neg/cv670_2666.txt',
'neg/cv671_5164.txt',
'neg/cv672_27988.txt',
'neg/cv673_25874.txt',
'neg/cv674_11593.txt',
'neg/cv675_22871.txt',
'neg/cv676_22202.txt',
'neg/cv677_18938.txt',
'neg/cv678_14887.txt',
'neg/cv679_28221.txt',
'neg/cv680_10533.txt',
'neg/cv681_9744.txt',
'neg/cv682_17947.txt',
'neg/cv683_13047.txt',
'neg/cv684_12727.txt',
```

```
'neg/cv685_5710.txt',
'neg/cv686_15553.txt',
'neg/cv687_22207.txt',
'neg/cv688_7884.txt',
'neg/cv689_13701.txt',
'neg/cv690_5425.txt',
'neg/cv691_5090.txt',
'neg/cv692_17026.txt',
'neg/cv693_19147.txt',
'neg/cv694_4526.txt',
'neg/cv695_22268.txt',
'neg/cv696_29619.txt',
'neg/cv697_12106.txt',
'neg/cv698_16930.txt',
'neg/cv699_7773.txt',
'neg/cv700_23163.txt',
'neg/cv701_15880.txt',
'neg/cv702_12371.txt',
'neg/cv703_17948.txt',
'neg/cv704_17622.txt',
'neg/cv705_11973.txt',
'neg/cv706_25883.txt',
'neg/cv707_11421.txt',
'neg/cv708_28539.txt',
'neg/cv709_11173.txt',
'neg/cv710_23745.txt',
'neg/cv711_12687.txt',
'neg/cv712_24217.txt',
'neg/cv713_29002.txt',
'neg/cv714_19704.txt',
'neg/cv715_19246.txt',
'neg/cv716_11153.txt',
'neg/cv717_17472.txt',
'neg/cv718_12227.txt',
'neg/cv719_5581.txt',
'neg/cv720_5383.txt',
'neg/cv721_28993.txt',
'neg/cv722_7571.txt',
'neg/cv723_9002.txt',
'neg/cv724_15265.txt',
'neg/cv725_10266.txt',
'neg/cv726_4365.txt',
'neg/cv727_5006.txt',
'neg/cv728_17931.txt',
'neg/cv729_10475.txt',
'neg/cv730_10729.txt',
'neg/cv731_3968.txt',
'neg/cv732_13092.txt',
```

```
'neg/cv733_9891.txt',
'neg/cv734_22821.txt',
'neg/cv735_20218.txt',
'neg/cv736_24947.txt',
'neg/cv737_28733.txt',
'neg/cv738_10287.txt',
'neg/cv739_12179.txt',
'neg/cv740_13643.txt',
'neg/cv741_12765.txt',
'neg/cv742_8279.txt',
'neg/cv743_17023.txt',
'neg/cv744_10091.txt',
'neg/cv745_14009.txt',
'neg/cv746_10471.txt',
'neg/cv747_18189.txt',
'neg/cv748_14044.txt',
'neg/cv749_18960.txt',
'neg/cv750_10606.txt',
'neg/cv751_17208.txt',
'neg/cv752_25330.txt',
'neg/cv753_11812.txt',
'neg/cv754_7709.txt',
'neg/cv755_24881.txt',
'neg/cv756_23676.txt',
'neg/cv757_10668.txt',
'neg/cv758_9740.txt',
'neg/cv759_15091.txt',
'neg/cv760_8977.txt',
'neg/cv761_13769.txt',
'neg/cv762_15604.txt',
'neg/cv763_16486.txt',
'neg/cv764_12701.txt',
'neg/cv765_20429.txt',
'neg/cv766_7983.txt',
'neg/cv767_15673.txt',
'neg/cv768_12709.txt',
'neg/cv769_8565.txt',
'neg/cv770_11061.txt',
'neg/cv771_28466.txt',
'neg/cv772_12971.txt',
'neg/cv773_20264.txt',
'neg/cv774_15488.txt',
'neg/cv775_17966.txt',
'neg/cv776_21934.txt',
'neg/cv777_10247.txt',
'neg/cv778_18629.txt',
'neg/cv779_18989.txt',
'neg/cv780_8467.txt',
```

```
'neg/cv781_5358.txt',
'neg/cv782_21078.txt',
'neg/cv783_14724.txt',
'neg/cv784_16077.txt',
'neg/cv785_23748.txt',
'neg/cv786_23608.txt',
'neg/cv787_15277.txt',
'neg/cv788_26409.txt',
'neg/cv789_12991.txt',
'neg/cv790_16202.txt',
'neg/cv791_17995.txt',
'neg/cv792_3257.txt',
'neg/cv793_15235.txt',
'neg/cv794_17353.txt',
'neg/cv795_10291.txt',
'neg/cv796_17243.txt',
'neg/cv797_7245.txt',
'neg/cv798_24779.txt',
'neg/cv799_19812.txt',
'neg/cv800_13494.txt',
'neg/cv801_26335.txt',
'neg/cv802_28381.txt',
'neg/cv803_8584.txt',
'neg/cv804_11763.txt',
'neg/cv805_21128.txt',
'neg/cv806_9405.txt',
'neg/cv807_23024.txt',
'neg/cv808_13773.txt',
'neg/cv809_5012.txt',
'neg/cv810_13660.txt',
'neg/cv811_22646.txt',
'neg/cv812_19051.txt',
'neg/cv813_6649.txt',
'neg/cv814_20316.txt',
'neg/cv815_23466.txt',
'neg/cv816_15257.txt',
'neg/cv817_3675.txt',
'neg/cv818_10698.txt',
'neg/cv819_9567.txt',
'neg/cv820_24157.txt',
'neg/cv821_29283.txt',
'neg/cv822_21545.txt',
'neg/cv823_17055.txt',
'neg/cv824_9335.txt',
'neg/cv825_5168.txt',
'neg/cv826_12761.txt',
'neg/cv827_19479.txt',
'neg/cv828_21392.txt',
```

```
'neg/cv829_21725.txt',
'neg/cv830_5778.txt',
'neg/cv831_16325.txt',
'neg/cv832_24713.txt',
'neg/cv833_11961.txt',
'neg/cv834_23192.txt',
'neg/cv835_20531.txt',
'neg/cv836_14311.txt',
'neg/cv837_27232.txt',
'neg/cv838_25886.txt',
'neg/cv839_22807.txt',
'neg/cv840_18033.txt',
'neg/cv841_3367.txt',
'neg/cv842_5702.txt',
'neg/cv843_17054.txt',
'neg/cv844_13890.txt',
'neg/cv845_15886.txt',
'neg/cv846_29359.txt',
'neg/cv847_20855.txt',
'neg/cv848_10061.txt',
'neg/cv849_17215.txt',
'neg/cv850_18185.txt',
'neg/cv851_21895.txt',
'neg/cv852_27512.txt',
'neg/cv853_29119.txt',
'neg/cv854_18955.txt',
'neg/cv855_22134.txt',
'neg/cv856_28882.txt',
'neg/cv857_17527.txt',
'neg/cv858_20266.txt',
'neg/cv859_15689.txt',
'neg/cv860_15520.txt',
'neg/cv861_12809.txt',
'neg/cv862_15924.txt',
'neg/cv863_7912.txt',
'neg/cv864_3087.txt',
'neg/cv865_28796.txt',
'neg/cv866_29447.txt',
'neg/cv867_18362.txt',
'neg/cv868_12799.txt',
'neg/cv869_24782.txt',
'neg/cv870_18090.txt',
'neg/cv871_25971.txt',
'neg/cv872_13710.txt',
'neg/cv873_19937.txt',
'neg/cv874_12182.txt',
'neg/cv875_5622.txt',
'neg/cv876_9633.txt',
```

```
'neg/cv877_29132.txt',
'neg/cv878_17204.txt',
'neg/cv879_16585.txt',
'neg/cv880_29629.txt',
'neg/cv881_14767.txt',
'neg/cv882_10042.txt',
'neg/cv883_27621.txt',
'neg/cv884_15230.txt',
'neg/cv885_13390.txt',
'neg/cv886_19210.txt',
'neg/cv887_5306.txt',
'neg/cv888_25678.txt',
'neg/cv889_22670.txt',
'neg/cv890_3515.txt',
'neg/cv891_6035.txt',
'neg/cv892_18788.txt',
'neg/cv893_26731.txt',
'neg/cv894_22140.txt',
'neg/cv895_22200.txt',
'neg/cv896_17819.txt',
'neg/cv897_11703.txt',
'neg/cv898_1576.txt',
'neg/cv899_17812.txt',
'neg/cv900_10800.txt',
'neg/cv901_11934.txt',
'neg/cv902_13217.txt',
'neg/cv903_18981.txt',
'neg/cv904_25663.txt',
'neg/cv905_28965.txt',
'neg/cv906_12332.txt',
'neg/cv907_3193.txt',
'neg/cv908_17779.txt',
'neg/cv909_9973.txt',
'neg/cv910_21930.txt',
'neg/cv911_21695.txt',
'neg/cv912_5562.txt',
'neg/cv913_29127.txt',
'neg/cv914_2856.txt',
'neg/cv915_9342.txt',
'neg/cv916_17034.txt',
'neg/cv917_29484.txt',
'neg/cv918_27080.txt',
'neg/cv919_18155.txt',
'neg/cv920_29423.txt',
'neg/cv921_13988.txt',
'neg/cv922_10185.txt',
'neg/cv923_11951.txt',
'neg/cv924_29397.txt',
```

```
'neg/cv925_9459.txt',
'neg/cv926_18471.txt',
'neg/cv927_11471.txt',
'neg/cv928_9478.txt',
'neg/cv929_1841.txt',
'neg/cv930_14949.txt',
'neg/cv931_18783.txt',
'neg/cv932_14854.txt',
'neg/cv933_24953.txt',
'neg/cv934_20426.txt',
'neg/cv935_24977.txt',
'neg/cv936_17473.txt',
'neg/cv937_9816.txt',
'neg/cv938_10706.txt',
'neg/cv939_11247.txt',
'neg/cv940_18935.txt',
'neg/cv941_10718.txt',
'neg/cv942_18509.txt',
'neg/cv943_23547.txt',
'neg/cv944_15042.txt',
'neg/cv945_13012.txt',
'neg/cv946_20084.txt',
'neg/cv947_11316.txt',
'neg/cv948_25870.txt',
'neg/cv949_21565.txt',
'neg/cv950_13478.txt',
'neg/cv951_11816.txt',
'neg/cv952_26375.txt',
'neg/cv953_7078.txt',
'neg/cv954_19932.txt',
'neg/cv955_26154.txt',
'neg/cv956_12547.txt',
'neg/cv957_9059.txt',
'neg/cv958_13020.txt',
'neg/cv959_16218.txt',
'neg/cv960_28877.txt',
'neg/cv961_5578.txt',
'neg/cv962_9813.txt',
'neg/cv963_7208.txt',
'neg/cv964_5794.txt',
'neg/cv965_26688.txt',
'neg/cv966_28671.txt',
'neg/cv967_5626.txt',
'neg/cv968_25413.txt',
'neg/cv969_14760.txt',
'neg/cv970_19532.txt',
'neg/cv971_11790.txt',
'neg/cv972_26837.txt',
```

```
 'neg/cv973_10171.txt',
 'neg/cv974_24303.txt',
 'neg/cv975_11920.txt',
 'neg/cv976_10724.txt',
 'neg/cv977_4776.txt',
 'neg/cv978_22192.txt',
 'neg/cv979_2029.txt',
 'neg/cv980_11851.txt',
 'neg/cv981_16679.txt',
 'neg/cv982_22209.txt',
 'neg/cv983_24219.txt',
 'neg/cv984_14006.txt',
 'neg/cv985_5964.txt',
 'neg/cv986_15092.txt',
 'neg/cv987_7394.txt',
 'neg/cv988_20168.txt',
 'neg/cv989_17297.txt',
 'neg/cv990_12443.txt',
 'neg/cv991_19973.txt',
 'neg/cv992_12806.txt',
 'neg/cv993_29565.txt',
 'neg/cv994_13229.txt',
 'neg/cv995_23113.txt',
 'neg/cv996_12447.txt',
 'neg/cv997_5152.txt',
 'neg/cv998_15691.txt',
 'neg/cv999_14636.txt',
 ...]
```

```
In [3]: len(movie_reviews.fileids())
```

```
Out[3]: 2000
```

Each corpus reader provides a variety of methods to read data from the corpus, depending on the format of the corpus. For example, plaintext corpora support methods to read the corpus as raw text, a list of words, a list of sentences, or a list of paragraphs.

```
In [4]: #movie_reviews.raw("neg/cv000_29416.txt").replace("\n", "").replace("'", '"').replace(
```

1.2 2. Building and testing the classifier

```
In [5]: from nltk.corpus import stopwords

        stops = stopwords.words('english')
        stops.extend('.,[,],(,),;,/,-,\',?,",:,<,>,n\'t,|,#,\'s,\",\'re,\'ve,\'ll,\'d,\'re'.sp
        stops.extend(',')
        stops
```

```
Out[5]: ['i',
         'me',
         'my',
         'myself',
         'we',
         'our',
         'ours',
         'ourselves',
         'you',
         "you're",
         "you've",
         "you'll",
         "you'd",
         'your',
         'yours',
         'yourself',
         'yourselves',
         'he',
         'him',
         'his',
         'himself',
         'she',
         "she's",
         'her',
         'hers',
         'herself',
         'it',
         "it's",
         'its',
         'itself',
         'they',
         'them',
         'their',
         'theirs',
         'themselves',
         'what',
         'which',
         'who',
         'whom',
         'this',
         'that',
         "that'll",
         'these',
         'those',
         'am',
         'is',
         'are',
         'was',
```

```
'were',
'be',
'been',
'being',
'have',
'has',
'had',
'having',
'do',
'does',
'did',
'doing',
'a',
'an',
'the',
'and',
'but',
'if',
'or',
'because',
'as',
'until',
'while',
'of',
'at',
'by',
'for',
'with',
'about',
'against',
'between',
'into',
'through',
'during',
'before',
'after',
'above',
'below',
'to',
'from',
'up',
'down',
'in',
'out',
'on',
'off',
'over',
'under',
```

```
'again',
'further',
'then',
'once',
'here',
'there',
'when',
'where',
'why',
'how',
'all',
'any',
'both',
'each',
'few',
'more',
'most',
'other',
'some',
'such',
'no',
'nor',
'not',
'only',
'own',
'same',
'so',
'than',
'too',
'very',
's',
't',
'can',
'will',
'just',
'don',
"don't",
'should',
"should've",
'now',
'd',
'll',
'm',
'o',
're',
've',
'y',
'ain',
```

```
'aren',
"aren't",
'couldn',
"couldn't",
'didn',
"didn't",
'doesn',
"doesn't",
'hadn',
"hadn't",
'hasn',
"hasn't",
'haven',
"haven't",
'isn',
"isn't",
'ma',
'mightn',
"mightn't",
'mustn',
"mustn't",
'needn',
"needn't",
'shan',
"shan't",
'shouldn',
"shouldn't",
'wasn',
"wasn't",
'weren',
"weren't",
'won',
"won't",
'wouldn',
"wouldn't",
'.',
'[',
']',
'(',
')',
';',
'/',
'-',
"'",
'?',
'"',
':',
'<',
```

```
 '>',
 "n't",
 '|',
 '#',
 "'s",
 '"',
 "'re",
 "'ve",
 "'ll",
 "'d",
 "'re",
 ',']
```

In [6]:
```
from nltk.classify import NaiveBayesClassifier
import nltk.classify.util # Utility functions and classes for classifiers. Contains fu

# Given a word, returns a dict {word: True}. This will be our feature in the classifie
def word_feats(words):
    return dict([(word, True) for word in words if word not in stops and word.isalpha()

pos_ids = movie_reviews.fileids('pos')
neg_ids = movie_reviews.fileids('neg')

len(pos_ids) + len(neg_ids)
```

Out[6]: 2000

In [7]:
```
# We take the positive/negative words, create the feature for such words, and store it
pos_feats = [(word_feats(movie_reviews.words(fileids=[f])), 'pos') for f in pos_ids]
neg_feats = [(word_feats(movie_reviews.words(fileids=[f])), 'neg') for f in neg_ids]

#pos_feats
```

In [8]:
```
# 3/4 of the features will be used for training.
pos_len_train = int(len(pos_feats) * 3 / 4)
neg_len_train = int(len(neg_feats) * 3 / 4)

pos_len_train
```

Out[8]: 750

In [9]:
```
train_feats = neg_feats[:neg_len_train] + pos_feats[:pos_len_train]
test_feats = neg_feats[neg_len_train:] + pos_feats[pos_len_train:]

# Training a NaiveBayesClassifier with our training feature words.
classifier = NaiveBayesClassifier.train(train_feats)

print('Accuracy: ', nltk.classify.util.accuracy(classifier, test_feats))
```

```
Accuracy:  0.712
```

```
In [10]: # We can see which words fit best in each class.
         classifier.show_most_informative_features()
```

```
Most Informative Features
             magnificent = True              pos : neg    =     15.0 : 1.0
             outstanding = True              pos : neg    =     13.6 : 1.0
               insulting = True              neg : pos    =     13.0 : 1.0
              vulnerable = True              pos : neg    =     12.3 : 1.0
               ludicrous = True              neg : pos    =     11.8 : 1.0
                  avoids = True              pos : neg    =     11.7 : 1.0
             uninvolving = True              neg : pos    =     11.7 : 1.0
              astounding = True              pos : neg    =     10.3 : 1.0
             fascination = True              pos : neg    =     10.3 : 1.0
                 idiotic = True              neg : pos    =      9.8 : 1.0
```

1.3 3. Classifying new data

```
In [11]: from nltk import word_tokenize, pos_tag

         sentence = "I feel so miserable, it makes me amazing"
         tokens = [word for word in word_tokenize(sentence) if word not in stops]
         tokens
```

```
Out[11]: ['I', 'feel', 'miserable', 'makes', 'amazing']
```

```
In [12]: feats = word_feats(word for word in tokens)
         feats
```

```
Out[12]: {'I': True, 'feel': True, 'miserable': True, 'makes': True, 'amazing': True}
```

```
In [13]: classifier.classify(feats)
```

```
Out[13]: 'pos'
```

```
In [14]: sentence2 = "You are a pathetic fool, a terrible excuse for a human being."
         tokens2 = [word for word in word_tokenize(sentence2) if word not in stops]
         tokens2
```

```
Out[14]: ['You', 'pathetic', 'fool', 'terrible', 'excuse', 'human']
```

```
In [15]: pos_tags2 = [pos for pos in pos_tag(tokens2) if pos[1] == 'JJ']
         pos_tags2
```

```
Out[15]: [('pathetic', 'JJ'), ('terrible', 'JJ')]
```

```
In [16]: feats2 = word_feats([word for (word,_) in pos_tags2])
         feats2
```

```
Out[16]: {'pathetic': True, 'terrible': True}

In [17]: classifier.classify(feats2)

Out[17]: 'neg'
```

1.4 4. Classifying News Documents in Categories (sport, humor, adventure, etc...)

```
In [18]: from nltk.corpus import brown
         from random import shuffle
         from nltk import NaiveBayesClassifier
         from nltk import FreqDist
         from nltk.classify import accuracy,apply_features
         from nltk.corpus import stopwords

         #given a document extract features (the presence or not of the best1500 top 1500 of f
         def document_features(doc):
             doc_set_words = set(doc)
             features_dic = {} #features is a dictionary
             for word in best1500_words_corpora:
                 features_dic['has(%s)' % word] = (word in doc_set_words)
             return features_dic

         #Word frecuency of words in corpora.
         #Return the top 1500 most used words.
         def init_corpora():
             words_in_corpora = FreqDist(w.lower() for w in brown.words())
             # Python 2
             #best1500 = words_in_corpora.keys()[:1500]
             # Python 3
             best1500 = list(words_in_corpora.keys())[:1500]

             return  nonstop(best1500)

         #Receives a list of words, and remove those in stopwords.
         def nonstop(listwords):
             return [word for word in listwords if word not in stopw]


         stopw = stopwords.words('english')

         best1500_words_corpora = init_corpora()
         documents = [( list(nonstop(brown.words(fileid)) ),categoria) for categoria in brown.
         shuffle(documents)
         print(len(documents))
         #print(documents)
         doc_features_set = [(document_features(d),c) for (d,c) in documents]
         categories = set([c for (_,c) in documents])
```

```
        print(categories)
        #print(len(doc_features_set))
```

```
500
{'religion', 'romance', 'adventure', 'learned', 'government', 'reviews', 'news', 'lore', 'edito
```

```
In [19]: train_set = doc_features_set[:350]
         #train_set = apply_features(document,features,doc_features_set[:200])
         classifier = NaiveBayesClassifier.train(train_set)
         classifier.show_most_informative_features(30)

         #We expect Adventura as 'cn02' contains words of a books of adventures
         print(classifier.classify(document_features(brown.words('cn02'))))
```

```
Most Informative Features
        has(television) = True             humor : learne =     24.9 : 1.0
              has(saw) = True            fictio : learne =     20.1 : 1.0
            has(quiet) = True            scienc : learne =     19.3 : 1.0
         has(afternoon) = True             humor : learne =     19.3 : 1.0
          has(midnight) = True            scienc : learne =     19.3 : 1.0
         has(announced) = True              news : learne =     19.3 : 1.0
              has(pay) = True            govern : learne =     19.3 : 1.0
            has(1961) = True              news : belles =     18.7 : 1.0
            has(funds) = True            govern : belles =     18.7 : 1.0
             has(wife) = True             humor : learne =     18.2 : 1.0
           has(church) = True            religi : learne =     18.1 : 1.0
          has(returned) = True            myster : learne =     18.0 : 1.0
             has(blue) = True            romanc : learne =     17.7 : 1.0
             has(race) = True            review : learne =     17.4 : 1.0
          has(audience) = True            review : learne =     17.4 : 1.0
              has(yes) = True            myster : learne =     17.4 : 1.0
           has(notice) = True            myster : belles =     17.3 : 1.0
           has(starts) = True            review : belles =     16.8 : 1.0
           has(fiscal) = True            govern : belles =     16.7 : 1.0
            has(board) = True            scienc : learne =     16.2 : 1.0
            has(night) = True            fictio : hobbie =     15.8 : 1.0
            has(stood) = True            myster : learne =     15.7 : 1.0
          has(schools) = True            editor : learne =     15.7 : 1.0
             has(says) = True            review : learne =     15.1 : 1.0
        has(telephone) = True             humor : learne =     14.9 : 1.0
           has(school) = True             humor : learne =     14.9 : 1.0
          has(meeting) = True             humor : belles =     14.4 : 1.0
              has(son) = True            romanc : learne =     14.3 : 1.0
             has(went) = False           govern : romanc =     14.1 : 1.0
         has(insisted) = True             humor : learne =     13.8 : 1.0
romance
```

```
In [20]: test_set  = doc_features_set[150:]
         print(accuracy(classifier,test_set))
```

```
0.7057142857142857
```