Text Classification

In [1]:

```
def gender_features(word):
    return{'last_letter':word[-1]}
```

In [2]:

```
gender_features('Obama')
```

Out[2]:

```
{'last_letter': 'a'}
```

In [3]:

```
gender_features('srinivas')
```

Out[3]:

```
{'last_letter': 's'}
```

In [43]:

```
gender_features('sharmila banu')
```

Out[43]:

```
{'last_letter': 'u'}
```

In [5]:

```
from nltk.corpus import names
```

In [12]:

```
names.words()
```

Out[12]:

```
['Abagael',
 'Abagail',
 'Abbe',
 'Abbey',
 'Abbi',
 'Abbie',
 'Abby',
 'Abigael',
 'Abigail',
 'Abigale',
 'Abra',
 'Acacia',
 'Ada',
 'Adah',
 'Adaline',
 'Adara',
 'Addie',
 'Addis',
 'Adel',
 'Adela',
 'Adelaide',
 'Adele',
 'Adelice',
 'Adelina',
 'Adelind',
```

'Adeline',
'Adella',
'Adelle',
'Adena',
'Adey',
'Adi',
'Adiana',
'Adina',
'Adora',
'Adore',
'Adoree',
'Adorne',
'Adrea',
'Adria',
'Adriaens',
'Adrian',
'Adriana',
'Adriane',
'Adrianna',
'Adrianne',
'Adrien',
'Adriena',
'Adrienne',
'Aeriel',
'Aeriela',
'Aeriell',
'Ag',
'Agace',
'Agata',
'Agatha',
'Agathe',
'Aggi',
'Aggie',
'Aggy',
'Agna',
'Agnella',
'Agnes',
'Agnese',
'Agnesse',
'Agneta',
'Agnola',
'Agretha',
'Aida',
'Aidan',
'Aigneis',
'Aila',
'Aile',
'Ailee',
'Aileen',
'Ailene',
'Ailey',
'Aili',
'Ailina',
'Ailyn',
'Aime',
'Aimee',
'Aimil',
'Aina',
'Aindrea',
'Ainslee',
'Ainsley',
'Ainslie',
'Ajay',
'Alaine',
'Alameda',
'Alana',
'Alanah',
'Alane',
'Alanna',
'Alayne',
'Alberta',
'Albertina',
'Albertine',
'Albina',
'Alecia',
'Aleda',
'Aleece',

```
'Aleecia',
'Aleen',
'Alejandra',
'Alejandrina',
'Alena',
'Alene',
'Alessandra',
'Aleta',
'Alethea',
'Alex',
'Alexa',
'Alexandra',
'Alexandrina',
'Alexi',
'Alexia',
'Alexina',
'Alexine',
'Alexis',
'Alfie',
'Alfreda',
'Ali',
'Alia',
'Alica',
'Alice',
'Alicea',
'Alicia',
'Alida',
'Alidia',
'Alina',
'Aline',
'Alis',
'Alisa',
'Alisha',
'Alison',
'Alissa',
'Alisun',
'Alix',
'Aliza',
'Alla',
'Alleen',
'Allegra',
'Allene',
'Alli',
'Allianora',
'Allie',
'Allina',
'Allis',
'Allison',
'Allissa',
'Allsun',
'Ally',
'Allyce',
'Allyn',
'Allys',
'Allyson',
'Alma',
'Almeda',
'Almeria',
'Almeta',
'Almira',
'Almire',
'Aloise',
'Aloisia',
'Aloysia',
'Alpa',
'Alta',
'Althea',
'Alvera',
'Alvina',
'Alvinia',
'Alvira',
'Alyce',
'Alyda',
'Alys',
'Alysa',
'Alyse',
'Alysia',
```

```
'Alyson',
'Alyss',
'Alyssa',
'Amabel',
'Amabelle',
'Amalea',
'Amalee',
'Amaleta',
'Amalia',
'Amalie',
'Amalita',
'Amalle',
'Amanda',
'Amandi',
'Amandie',
'Amandy',
'Amara',
'Amargo',
'Amata',
'Amber',
'Amberly',
'Ambrosia',
'Ambur',
'Ame',
'Amelia',
'Amelie',
'Amelina',
'Ameline',
'Amelita',
'Ami',
'Amie',
'Amity',
'Ammamaria',
'Amy',
'Ana',
'Anabel',
'Anabella',
'Anabelle',
'Anais',
'Analiese',
'Analise',
'Anallese',
'Anallise',
'Anastasia',
'Anastasie',
'Anastassia',
'Anatola',
'Andee',
'Andi',
'Andie',
'Andra',
'Andrea',
'Andreana',
'Andree',
'Andrei',
'Andria',
'Andriana',
'Andriette',
'Andromache',
'Andromeda',
'Andy',
'Anestassia',
'Anet',
'Anett',
'Anetta',
'Anette',
'Ange',
'Angel',
'Angela',
'Angele',
'Angelia',
'Angelica',
'Angelika',
'Angelina',
'Angeline',
'Angelique',
'Angelita',
```

```
'Angelle',
'Angie',
'Angil',
'Angy',
'Ania',
'Anica',
'Anissa',
'Anita',
'Anitra',
'Anja',
'Anjanette',
'Anjela',
'Ann',
'Ann-Mari',
'Ann-Marie',
'Anna',
'Anna-Diana',
'Anna-Diane',
'Anna-Maria',
'Annabal',
'Annabel',
'Annabela',
'Annabell',
'Annabella',
'Annabelle',
'Annadiana',
'Annadiane',
'Annalee',
'Annalena',
'Annaliese',
'Annalisa',
'Annalise',
'Annalyse',
'Annamari',
'Annamaria',
'Annamarie',
'Anne',
'Anne-Corinne',
'Anne-Mar',
'Anne-Marie',
'Annecorinne',
'Anneliese',
'Annelise',
'Annemarie',
'Annetta',
'Annette',
'Anni',
'Annice',
'Annie',
'Annissa',
'Annmaria',
'Annmarie',
'Annnora',
'Annora',
'Anny',
'Anselma',
'Ansley',
'Anstice',
'Anthe',
'Anthea',
'Anthia',
'Antoinette',
'Antonella',
'Antonetta',
'Antonia',
'Antonie',
'Antonietta',
'Antonina',
'Anya',
'Aphrodite',
'Appolonia',
'April',
'Aprilette',
'Ara',
'Arabel',
'Arabela',
'Arabele',
```

```
'Arabella',
'Arabelle',
'Arda',
'Ardath',
'Ardeen',
'Ardelia',
'Ardelis',
'Ardella',
'Ardelle',
'Arden',
'Ardene',
'Ardenia',
'Ardine',
'Ardis',
'Ardith',
'Ardra',
'Ardyce',
'Ardys',
'Ardyth',
'Aretha',
'Ariadne',
'Ariana',
'Arianne',
'Aridatha',
'Ariel',
'Ariela',
'Ariella',
'Arielle',
'Arlana',
'Arlee',
'Arleen',
'Arlen',
'Arlena',
'Arlene',
'Arleta',
'Arlette',
'Arleyne',
'Arlie',
'Arliene',
'Arlina',
'Arlinda',
'Arline',
'Arly',
'Arlyn',
'Arlyne',
'Aryn',
'Ashely',
'Ashlee',
'Ashleigh',
'Ashlen',
'Ashley',
'Ashli',
'Ashlie',
'Ashly',
'Asia',
'Astra',
'Astrid',
'Astrix',
'Atalanta',
'Athena',
'Athene',
'Atlanta',
'Atlante',
'Auberta',
'Aubine',
'Aubree',
'Aubrette',
'Aubrey',
'Aubrie',
'Aubry',
'Audi',
'Audie',
'Audra',
'Audre',
'Audrey',
'Audrie',
'Audry',
```

```
'Audrye',
'Audy',
'Augusta',
'Auguste',
'Augustina',
'Augustine',
'Aura',
'Aurea',
'Aurel',
'Aurelea',
'Aurelia',
'Aurelie',
'Auria',
'Aurie',
'Aurilia',
'Aurlie',
'Auroora',
'Aurora',
'Aurore',
'Austin',
'Austina',
'Austine',
'Ava',
'Aveline',
'Averil',
'Averyl',
'Avie',
'Avis',
'Aviva',
'Avivah',
'Avril',
'Avrit',
'Ayn',
'Bab',
'Babara',
'Babette',
'Babita',
'Babs',
'Bambi',
'Bambie',
'Bamby',
'Barb',
'Barbabra',
'Barbara',
'Barbara-Anne',
'Barbaraanne',
'Barbe',
'Barbee',
'Barbette',
'Barbey',
'Barbi',
'Barbie',
'Barbra',
'Barby',
'Bari',
'Barrie',
'Barry',
'Basia',
'Bathsheba',
'Batsheva',
'Bea',
'Beatrice',
'Beatrisa',
'Beatrix',
'Beatriz',
'Beau',
'Bebe',
'Becca',
'Becka',
'Becki',
'Beckie',
'Becky',
'Bee',
'Beilul',
'Beitris',
'Bekki',
'Bel',
```

```
'Belia',
'Belicia',
'Belinda',
'Belita',
'Bell',
'Bella',
'Bellamy',
'Bellanca',
'Belle',
'Bellina',
'Belva',
'Belvia',
'Bendite',
'Benedetta',
'Benedicta',
'Benedikta',
'Benetta',
'Benita',
'Benni',
'Bennie',
'Benny',
'Benoite',
'Berenice',
'Beret',
'Berget',
'Berna',
'Bernadene',
'Bernadette',
'Bernadina',
'Bernadine',
'Bernardina',
'Bernardine',
'Bernelle',
'Bernete',
'Bernetta',
'Bernette',
'Berni',
'Bernice',
'Bernie',
'Bernita',
'Berny',
'Berri',
'Berrie',
'Berry',
'Bert',
'Berta',
'Berte',
'Bertha',
'Berthe',
'Berti',
'Bertie',
'Bertina',
'Bertine',
'Berty',
'Beryl',
'Beryle',
'Bess',
'Bessie',
'Bessy',
'Beth',
'Bethanne',
'Bethany',
'Bethena',
'Bethina',
'Betsey',
'Betsy',
'Betta',
'Bette',
'Bette-Ann',
'Betteann',
'Betteanne',
'Betti',
'Bettie',
'Bettina',
'Bettine',
'Betty',
'Bettye',
```

```
'Beulah',
'Bev',
'Beverie',
'Beverlee',
'Beverlie',
'Beverly',
'Bevvy',
'Bianca',
'Bianka',
'Biddy',
'Bidget',
'Bill',
'Billi',
'Billie',
'Billy',
'Binni',
'Binnie',
'Binny',
'Bird',
'Birdie',
'Birgit',
'Birgitta',
'Blair',
'Blaire',
'Blake',
'Blakelee',
'Blakeley',
'Blanca',
'Blanch',
'Blancha',
'Blanche',
'Blinni',
'Blinnie',
'Blinny',
'Bliss',
'Blisse',
'Blithe',
'Blondell',
'Blondelle',
'Blondie',
'Blondy',
'Blythe',
'Bo',
'Bobbette',
'Bobbi',
'Bobbie',
'Bobby',
'Bobette',
'Bobina',
'Bobine',
'Bobinette',
'Bonita',
'Bonnee',
'Bonni',
'Bonnie',
'Bonny',
'Brana',
'Brandais',
'Brande',
'Brandea',
'Brandi',
'Brandice',
'Brandie',
'Brandise',
'Brandy',
'Brea',
'Breanne',
'Brear',
'Bree',
'Breena',
'Bren',
'Brena',
'Brenda',
'Brenn',
'Brenna',
'Brett',
'Bria',
```

```
'Briana',
'Brianna',
'Brianne',
'Bride',
'Bridget',
'Bridgett',
'Bridgette',
'Bridie',
'Brier',
'Brietta',
'Brigid',
'Brigida',
'Brigit',
'Brigitta',
'Brigitte',
'Brina',
'Briney',
'Briny',
'Brit',
'Brita',
'Britaney',
'Britani',
'Briteny',
'Britney',
'Britni',
'Britt',
'Britta',
'Brittan',
'Brittany',
'Britte',
'Brittney',
'Brook',
'Brooke',
'Brooks',
'Brunella',
'Brunhilda',
'Brunhilde',
'Bryana',
'Bryn',
'Bryna',
'Brynn',
'Brynna',
'Brynne',
'Buffy',
'Bunni',
'Bunnie',
'Bunny',
'Burta',
'Cabrina',
'Cacilia',
'Cacilie',
'Caitlin',
'Caitrin',
'Cal',
'Calida',
'Calla',
'Calley',
'Calli',
'Callida',
'Callie',
'Cally',
'Calypso',
'Cam',
'Camala',
'Camel',
'Camella',
'Camellia',
'Cameo',
'Cami',
'Camila',
'Camile',
'Camilla',
'Camille',
'Cammi',
'Cammie',
'Cammy',
'Canada',
```

```
'Candace',
'Candi',
'Candice',
'Candida',
'Candide',
'Candie',
'Candis',
'Candra',
'Candy',
'Cappella',
'Caprice',
'Cara',
'Caralie',
'Caren',
'Carena',
'Caresa',
'Caressa',
'Caresse',
'Carey',
'Cari',
'Caria',
'Carie',
'Caril',
'Carilyn',
'Carin',
'Carina',
'Carine',
'Cariotta',
'Carissa',
'Carita',
'Caritta',
'Carla',
'Carlee',
'Carleen',
'Carlen',
'Carlena',
'Carlene',
'Carley',
'Carli',
'Carlie',
'Carlin',
'Carlina',
'Carline',
'Carlisle',
'Carlita',
'Carlota',
'Carlotta',
'Carly',
'Carlye',
'Carlyn',
'Carlynn',
'Carlynne',
'Carma',
'Carmel',
'Carmela',
'Carmelia',
'Carmelina',
'Carmelita',
'Carmella',
'Carmelle',
'Carmen',
'Carmina',
'Carmine',
'Carmita',
'Carmon',
'Caro',
'Carol',
'Carol-Jean',
'Carola',
'Carolan',
'Carolann',
'Carole',
'Carolee',
'Caroleen',
'Carolie',
'Carolin',
'Carolina',
```

'Caroline',
'Caroljean',
'Carolyn',
'Carolyne',
'Carolynn',
'Caron',
'Carree',
'Carri',
'Carrie',
'Carrissa',
'Carrol',
'Carroll',
'Carry',
'Cary',
'Caryl',
'Caryn',
'Casandra',
'Casey',
'Casi',
'Casia',
'Casie',
'Cass',
'Cassandra',
'Cassandre',
'Cassandry',
'Cassaundra',
'Cassey',
'Cassi',
'Cassie',
'Cassondra',
'Cassy',
'Cat',
'Catarina',
'Cate',
'Caterina',
'Catha',
'Catharina',
'Catharine',
'Cathe',
'Cathee',
'Catherin',
'Catherina',
'Catherine',
'Cathi',
'Cathie',
'Cathleen',
'Cathlene',
'Cathrin',
'Cathrine',
'Cathryn',
'Cathy',
'Cathyleen',
'Cati',
'Catie',
'Catina',
'Catlaina',
'Catlee',
'Catlin',
'Catrina',
'Catriona',
'Caty',
'Cayla',
'Cecelia',
'Cecil',
'Cecile',
'Ceciley',
'Cecilia',
'Cecilla',
'Cecily',
'Ceil',
'Cele',
'Celene',
'Celesta',
'Celeste',
'Celestia',
'Celestina',
'Celestine',

'Celestyn',
'Celestyna',
'Celia',
'Celie',
'Celina',
'Celinda',
'Celine',
'Celinka',
'Celisse',
'Celle',
'Cesya',
'Chad',
'Chanda',
'Chandal',
'Chandra',
'Channa',
'Chantal',
'Chantalle',
'Charil',
'Charin',
'Charis',
'Charissa',
'Charisse',
'Charita',
'Charity',
'Charla',
'Charlean',
'Charleen',
'Charlena',
'Charlene',
'Charline',
'Charlot',
'Charlott',
'Charlotta',
'Charlotte',
'Charmain',
'Charmaine',
'Charmane',
'Charmian',
'Charmine',
'Charmion',
'Charo',
'Charyl',
'Chastity',
'Chelsae',
'Chelsea',
'Chelsey',
'Chelsie',
'Chelsy',
'Cher',
'Chere',
'Cherey',
'Cheri',
'Cherianne',
'Cherice',
'Cherida',
'Cherie',
'Cherilyn',
'Cherilynn',
'Cherin',
'Cherise',
'Cherish',
'Cherlyn',
'Cherri',
'Cherrita',
'Cherry',
'Chery',
'Cherye',
'Cheryl',
'Cheslie',
'Chiarra',
'Chickie',
'Chicky',
'Chiquita',
'Chloe',
'Chloette',
'Chloris',

```
 'Chris',
 'Chriss',
 'Chrissa',
 'Chrissie',
 'Chrissy',
 'Christa',
 'Christabel',
 'Christabella',
 'Christabelle',
 'Christal',
 'Christalle',
 'Christan',
 'Christean',
 'Christel',
 'Christen',
 'Christi',
 'Christian',
 'Christiana',
 'Christiane',
 'Christie',
 'Christin',
 'Christina',
 'Christine',
 'Christy',
 'Christyna',
 'Chrysa',
 'Chrysler',
 'Chrystal',
 'Chryste',
 'Chrystel',
 'Ciara',
 'Cicely',
 'Cicily',
 'Ciel',
 'Cilka',
 'Cinda',
 'Cindee',
 'Cindelyn',
 'Cinderella',
 'Cindi',
 'Cindie',
 'Cindra',
 'Cindy',
 'Cinnamon',
 'Cissie',
 'Cissy',
 'Clair',
 'Claire',
 'Clara',
 'Clarabelle',
 'Clare',
 ...]
```

In [13]:

```
print(len(names.words()))
```

```
7944
```

In [10]:

```
labeled_names =([(name,'male')for name in names.words('male.txt')]+[(name,'female')for name in
names.words('female.txt')])
```

Tagging the Gender

In [22]:

```
import random
random.shuffle(labeled_names)
```

In [23]:

```
featuresets = [(gender_features(n),gender)for (n,gender)in labeled_names]
```

In [24]:

```
train_set,test_set = featuresets[5000:],featuresets[:2000]
```

In [26]:

```
import nltk
classifier = nltk.NaiveBayesClassifier.train(train_set)
classifier.show_most_informative_features(10)
```

```
Most Informative Features
             last_letter = 'k'              male : female =     36.8 : 1.0
             last_letter = 'a'            female : male   =     33.6 : 1.0
             last_letter = 'm'              male : female =     16.0 : 1.0
             last_letter = 'r'              male : female =      8.5 : 1.0
             last_letter = 'o'              male : female =      7.8 : 1.0
             last_letter = 'z'              male : female =      7.4 : 1.0
             last_letter = 'p'              male : female =      7.4 : 1.0
             last_letter = 'g'              male : female =      7.0 : 1.0
             last_letter = 't'              male : female =      6.1 : 1.0
             last_letter = 'd'              male : female =      5.8 : 1.0
```

In [27]:

```
classifier.classify(gender_features('srija'))
```

Out[27]:

```
'female'
```

In [29]:

```
classifier.classify(gender_features('srija Parimi'))
```

Out[29]:

```
'female'
```

In [45]:

```
classifier.classify(gender_features('Sharmila'))
```

Out[45]:

```
'female'
```

In [46]:

```
classifier.classify(gender_features('Sharmila Banu'))
```

Out[46]:

```
'male'
```

In [28]:

```
nltk.classify.accuracy(classifier,test_set)
```

Out[28]:

```
0.7655
```

Pos Tagging

In [33]:

```
import nltk
tokens = nltk.word_tokenize("hello everyone have a nice day")
print("Parts of Speech:",nltk.pos_tag(tokens))
```

```
Parts of Speech: [('hello', 'NN'), ('everyone', 'NN'), ('have', 'VBP'), ('a', 'DT'), ('nice', 'JJ'
), ('day', 'NN')]
```

In [30]:

```
import nltk
from nltk.corpus import stopwords
```

In [31]:

```
from nltk.tokenize import word_tokenize, sent_tokenize
stop_words = set(stopwords.words('english'))
```

In [36]:

```
txt = """ It began derailing after season one. His world was FUBAR by then. A promising young
teammate, Richard Hamilton, had dared to stand up to him in a mutual searing of egos, and found hi
mself traded. The mounting dissension on the team called to mind a word that Michael Jordan and so
me of his old Chicago Bulls associates exchanged during the Bulls' glory days to describe somethin
g or someone gone bad indefinitely. It was a code word, an acronym. FUBAR: Fucked Up Beyond All Re
cognition. By his last season, the Washington Wizards were hopelessly FUBAR.

Michael Jordan's three years in Washington -- about a year and nine months as an official executiv
e and two seasons as a player -- were troubled from the start. Before his comeback began, The Wash
ington Post dispatched me to watch him for an entire season, and much of a second. I valued the ex
perience, even the awfulness, which I hesitate admitting because I realize it sounds peculiar and
a little perverse. But if you wanted to know what forces -- money and a sense of entitlement, most
of all -- coarsened professional sports in the last gasps of the 20th century and the beginning of
the new millennium, it behooved you to have been witness to the Wizards and the Michael Show."""
```

In [37]:

```
tokenized = sent_tokenize(txt)
for i in tokenized:
    wordsList = nltk.word_tokenize(i)
```

In [38]:

```
wordsList = [w for w in wordsList if not w in stop_words]
```

In [41]:

```
tagged = nltk.pos_tag(wordsList)
```

In [42]:

```
  print(tagged)
```

```
[('But', 'CC'), ('wanted', 'VBD'), ('know', 'JJ'), ('forces', 'NNS'), ('--', ':'), ('money',
'NN'), ('sense', 'NN'), ('entitlement', 'NN'), (',', ','), ('--', ':'), ('coarsened', 'VBD'),
('professional', 'JJ'), ('sports', 'NNS'), ('last', 'JJ'), ('gasps', 'JJ'), ('20th', 'JJ'), ('cent
ury', 'NN'), ('beginning', 'VBG'), ('new', 'JJ'), ('millennium', 'NN'), (',', ','), ('behooved', '
VBD'), ('witness', 'JJ'), ('Wizards', 'IN'), ('Michael', 'NNP'), ('Show', 'NNP'), ('.', '.')]
```

In [ ]: