Creating features from text

1. Looking at the character sets: Latin, Cyrillic, Greek, CJK (Chinese, Japanese, Korean), etc. can provide a first categorization into language families.
2. Looking at one-letter, two-letter or three-letter words and their frequency in a text.
3. Character n-grams and their frequency.
4. Word n-grams and their frequency.

What are n-grams?

Word n-grams deal with sentences. “I like red cherries” will have as bigrams: “I like”, “like red”, “red cherries”. N-grams are a common model for representing language in the field of Natural Language Processing (a subfield of Artificial Intelligence).
Comparing character bigrams in different languages

Top 30 bigrams for English (%).

<table>
<thead>
<tr>
<th>Bigram</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TH</td>
<td>2.71</td>
</tr>
<tr>
<td>HE</td>
<td>2.33</td>
</tr>
<tr>
<td>IN</td>
<td>2.03</td>
</tr>
<tr>
<td>ER</td>
<td>1.78</td>
</tr>
<tr>
<td>AN</td>
<td>1.61</td>
</tr>
<tr>
<td>RE</td>
<td>1.41</td>
</tr>
<tr>
<td>ES</td>
<td>1.32</td>
</tr>
<tr>
<td>ON</td>
<td>1.32</td>
</tr>
<tr>
<td>ST</td>
<td>1.25</td>
</tr>
<tr>
<td>NT</td>
<td>1.17</td>
</tr>
<tr>
<td>DE</td>
<td>2.57</td>
</tr>
<tr>
<td>EN</td>
<td>2.31</td>
</tr>
<tr>
<td>EN</td>
<td>2.27</td>
</tr>
<tr>
<td>EL</td>
<td>2.01</td>
</tr>
<tr>
<td>LA</td>
<td>1.80</td>
</tr>
<tr>
<td>OS</td>
<td>1.79</td>
</tr>
<tr>
<td>OR</td>
<td>1.61</td>
</tr>
<tr>
<td>AS</td>
<td>1.56</td>
</tr>
<tr>
<td>ER</td>
<td>1.52</td>
</tr>
<tr>
<td>RA</td>
<td>1.47</td>
</tr>
<tr>
<td>EN</td>
<td>2.27</td>
</tr>
<tr>
<td>EA</td>
<td>1.00</td>
</tr>
<tr>
<td>TE</td>
<td>1.00</td>
</tr>
<tr>
<td>TO</td>
<td>1.07</td>
</tr>
<tr>
<td>TI</td>
<td>1.06</td>
</tr>
<tr>
<td>TH</td>
<td>1.07</td>
</tr>
<tr>
<td>TH</td>
<td>1.06</td>
</tr>
<tr>
<td>AH</td>
<td>0.98</td>
</tr>
<tr>
<td>OU</td>
<td>0.98</td>
</tr>
<tr>
<td>EN</td>
<td>0.98</td>
</tr>
<tr>
<td>ST</td>
<td>0.98</td>
</tr>
<tr>
<td>RA</td>
<td>0.98</td>
</tr>
<tr>
<td>EN</td>
<td>0.98</td>
</tr>
<tr>
<td>EN</td>
<td>0.98</td>
</tr>
<tr>
<td>EN</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Top 30 bigrams for Spanish (%).

<table>
<thead>
<tr>
<th>Bigram</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>1.43</td>
</tr>
<tr>
<td>EN</td>
<td>1.43</td>
</tr>
<tr>
<td>EN</td>
<td>1.42</td>
</tr>
<tr>
<td>EN</td>
<td>1.33</td>
</tr>
<tr>
<td>LA</td>
<td>1.24</td>
</tr>
<tr>
<td>OS</td>
<td>1.22</td>
</tr>
<tr>
<td>OR</td>
<td>1.21</td>
</tr>
<tr>
<td>AS</td>
<td>1.15</td>
</tr>
<tr>
<td>ER</td>
<td>1.13</td>
</tr>
<tr>
<td>RA</td>
<td>1.11</td>
</tr>
</tbody>
</table>

To notice:

- The top 3 bigrams for English cannot be found at all in the list of Spanish bigrams.
- The two lists have 14 bigrams in common out of 30 (less than half).
- The bigrams that are in common have different frequency. E.g., EN is 2.27 in Spanish and 1.13 in English.

Note: These bigrams were calculated from a large set of news stories. Because the word “the” is the most common word in English speech, that explains why the two bigrams “th” and “he” are at the top. If we use only the vocabulary of English words, the list will change. The most common bigram becomes “in”, because of the many words that start with “in” or that end in “ing”.

English word length distribution

Problem: Given the dictionary of all English words, what is the distribution of words by length?

Solution 1 (requires two separate loops)
1. Iterate over the list of words to find the length of each word and store it into a new list. [Accumulation in a list via mapping]
2. Iterate over the list of lengths and store it into a dictionary to keep track of the number of times we encounter each length. [Accumulation via a dictionary]

Solution 2 (requires one loop)
1. Iterate over the list of words to find the length of each word and immediately store it into a dictionary. [Accumulation via a dictionary]

English word length distribution - Code

Solution 1 (separate accumulation in two steps)

Step 1
```python
lengthsList = [len(word) for word in englishwords]
```

Step 2
```python
lengthsDct = {}
for length in lengthsList:
    lengthsDct[length] = lengthsDct.get(length, 0) + 1
```

Solution 2 (one single loop accumulating into the dictionary)
```python
lengthsDct2 = {}
for word in englishwords:
    length = len(word)
    lengthsDct2[length] = lengthsDct2.get(length, 0) + 1
```
Fun with statistics

In Statistics, it is common to describe a dataset (e.g., the list of the lengths of all English words) in terms of descriptive statistics: the mean, the median, the mode (the value that occurs the most), the variance, the standard deviation, etc. All these statistics can be calculated with the operations we have been learning.

- The **mean** is the sum of all list elements divided by the length of the list. (sum = accumulation to a number)
- The **median** is the middle element of a sorted list. (Sorting will be covered in Lec 16.)
- The **mode** is the most frequent element (i.e., the max value from the frequency dictionary.)
- The **variance** is the sum of the squares of the difference of each item to the mean.
- The **standard deviation** is the square of the variance.

Try it out

Using the `lengthsList` and `lengthsDct` from the previous slide, you can practice calculating these statistics with Python code.

You should find that both the median and the mode are 8.

Building character n-grams

Unigrams: 
```python
word = 'boston'
list(word)
```

Bigrams: 
```python
'bo', 'os', 'st', 'to', 'on'
```

Trigrams: 
```python
'bos', 'ost', 'sto', 'ton'
```

N-gram frequency distributions

- There are 66230 words in `englishwords`, we want to avoid iterating over them many times to create all n-gram distributions.
- We can create all three n-gram distributions in one single loop.
- Imagine we have three functions: `unigrams`, `bigrams`, `trigrams` that contain as bodies the statements in 16-10.
- Imagine also a function `storeNgrams` that takes a list of n-grams and a dictionary and adds the list elements to it, by updating their count.

```python
unigramsDct = {}
bigramsDct = {}
trigramsDct = {}

for word in englishwords:
    bigramsDct[word] = bigramsDct.get(word, 0) + 1
    bigramsDct = bigramsDct
```

Concepts in this slide:

- How to avoid multiple iterations by creating helper functions?
Mutating Dictionaries via aliasing

A dictionary can be mutated via aliasing.

```python
def storeNgrams(ngramsList, ngramsDict):
    """Given a list of items and a dictionary, update the counts of the dictionary keys."
    for ngram in ngramsList:
        ngramsDict[ngram] = ngramsDict.get(ngram, 0) + 1
```

Function Call Frames

storeNgrams(ngrams1, unigramsDct) → unigramsDct = 'b' 1
ngramsList  ngramsDict

storeNgrams(ngrams2, bigramsDct) → bigramsDct = 'bo' 1
ngramsList  ngramsDict

storeNgrams(ngrams3, trigramsDct) → trigramsDct = 'bos' 1
ngramsList  ngramsDict

Analyzing the Results

- Predict what will be the max lengths for the unigramsDct, bigramsDct, and trigramsDct: ____, ____, ____
- Do you expect that all dictionaries will have that max length? Explain.
- Predict the top 3 unigrams, top 3 bigrams, and top 3 unigrams.
- How to write a function sortItemsInFreqDict that given a frequency
dictionary will return the sorted list (in descending order) of its items, based
on the value of each (key/value) item? Hint: use lambda.
(Note: In Fall 2018, we will not cover sorting until Lec 16. So come back to
this question after Lec 16.)
- Which will be more frequent (have the highest values): the top unigrams, the
top bigrams, or the top trigrams?

Dictionary Comprehension

Very much like list comprehension: use {} instead of [] and create
pairs with the colon syntax, e.g., aKey: aValue.

Syntax:  
        { aKey: aValue for aKey in sequence}

Example: Write a dictionary comprehension that pairs words with their
lengths.

```python
In [1]: wordsLst = 'the autumn is dragging its feet'.split()
```

```python
In [2]: {word: len(word) for word in wordsLst}
```

Out[2]: {'autumn': 6, 'dragging': 8, 'feet': 4, 'is': 2, 'its': 3, 'the': 3}

Important

We can use dictionary comprehension in situations when we want to start accumulation
with a complex data structure (see next slide).

Accumulating in a dictionary of dictionaries

Problem: How can we create a dictionary that has
two level of keys? In the first level, each key is a
unigram, in the second level the keys are bigrams that
start with the unigram. [See example on the right.]

Solution 1: Assume we already have bigramsDict:

```python
{ 'a': { 'aa': 19, 'ab': 1665, 'ac': 2387, 'ad': 1685, ... },
  'b': { 'ba': 1431, 'bb': 417, 'bc': 25, 'bd': 35, ... },
  ...
}
```

```python
from string import lowercase  # 'abcdefghijklmnopqrstuvwxyz'
# create the dict with unigrams as keys and empty dict as values
bigramsByFirstLetter = {char: {} for char in lowercase}
```

```python
for bigram in bigramsDct:
    unigram = bigram[0]
    # assign the second level of keys
    bigramsByFirstLetter[unigram][bigram] = bigramsDct[bigram]```
Solution 2: We don’t have bigrams, we create them as we iterate over the list of words.

```python
from string import lowercase  # 'abcdefghijklmnopqrstuvwxyz'

# create the dict with unigrams as keys and empty dict as values
bigramsByFirstLetter = {char: {} for char in lowercase}

for word in englishwords:
    # create list of bigrams from word
    bigramsList = bigrams(word)
    # iterate over bigrams
    for bigram in bigramsList:
        unigram = bigram[0]
        # access the nested bigram dict for easy reference
        bigramsDct = bigramsByFirstLetter[unigram]
        # increase frequency counter
        bigramsDct[bigram] = bigramsDct.get(bigram, 0) + 1
```

**New collection type: set**

Sets are collections that have a unique number of items. They are useful to find the number of unique values in lists.

Find unigrams in a phrase:

```python
set([char for char in list(phrase)])
```

**Summary**

1. Lists and dictionaries are powerful data structures that are used routinely to perform complex data analysis tasks such as transforming data from one form to another.
2. Accumulation is a very common pattern in problem solving: we accumulate frequencies (counts) as we encounter new data; or we organize data as nested dictionaries of dictionaries or dictionaries of lists.
3. When we need to accumulate into nested structures, first always draw a picture of what the structure would look like, in order to visualize what needs to be created through code.
4. Dictionaries are mutable and they can be changed via aliasing (two different variables pointing to the same dictionary object).
5. The nested structures would need double subscripting operations (e.g., see last statement in slide 16-16). If this is conceptually difficult, you can store the inner structure into a temporary variable and work with that instead. Because of aliasing, this temporary variable will be directly mutating the entire dictionary. [See second from last statement in 16-17.]
6. Use a dictionary comprehension statement whenever you need to create a dictionary of dictionaries or a dictionary of lists in the case that the keys of the outer dictionary are known.
7. `set` is a new collection type. To create a set, always use the syntax `set({})`. Sets are also mutable, you can grow them by using the method `add`. 

```python
# add for sets is like append for lists
empty.add(3)
```