Cultural Analytics

ENGL 64.05

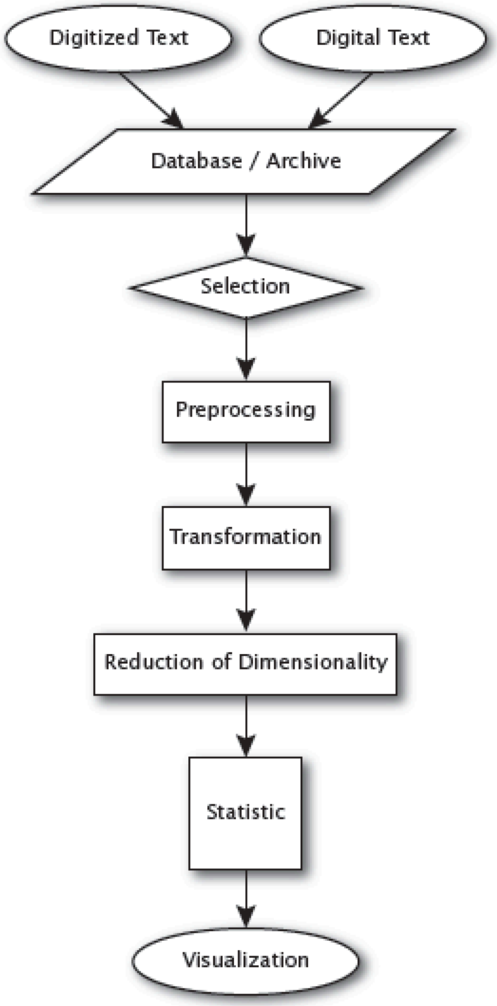
Fall 2019

Prof. James E. Dobson

October 4, 2019

Measuring Similarity

- **Digitized Texts**: Which editions? What to include within the text (preface? title page? paratexts?)
- **Database/Archive**: Why these texts? Are they already assumed to mean something within this archive? Is it representative?
- **Selection**: Which subset? Why? Is it representative?
- **Preprocessing**: Case-insensitive? Dropping or converting accented terms? Preserving Stopwords?
- **Transformation**: Conversion into vector space as bigrams or single terms?
- **Reduction of Dimensionality**: How mean features? Maximum document frequency? Minimum document frequency?
- **Statistic**: Which distance metric? Why this one?
- **Visualization**: How do we present the distance matrix?



Cultural Analytics

First Ten Features / First Ten Books

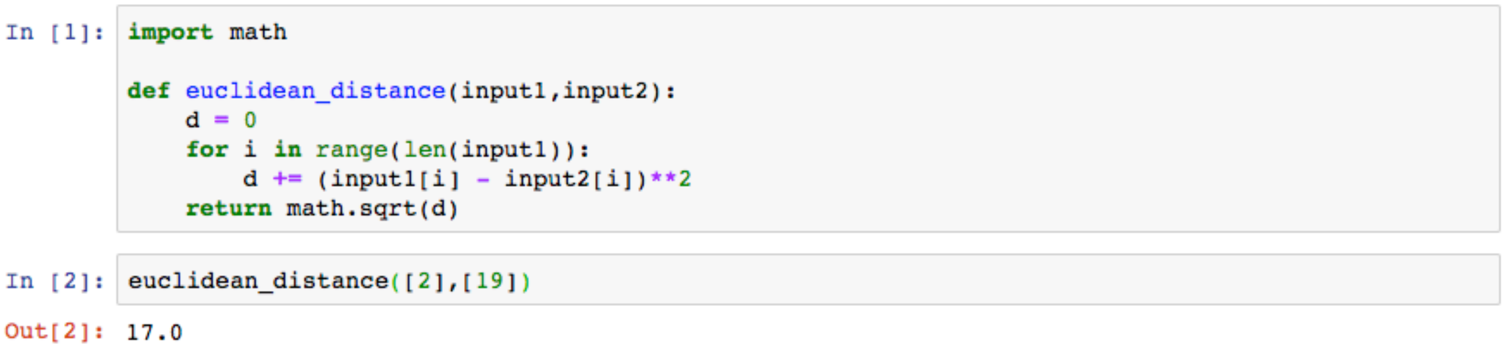
```
[0  0  0  0  0  0  0  0  0  0]
[0  22  0  0  0  0  0  0  0  0]
[0  0  0  0  0  0  0  0  0  0]
[0  0  0  0  0  0  0  0  0  0]
[0  12  0  0  0  0  0  0  0  0]
[27  1  0  0  0  0  0  0  0  0]
[0  0  0  0  0  0  0  0  0  0]
[0  2  0  0  0  0  0  0  0  0]
[0  0  0  0  0  0  0  0  0  0]
```

Euclidean Distance

def: The shortest straight line between two points (ruler distance). The square root of the summed squared differences.

2 19

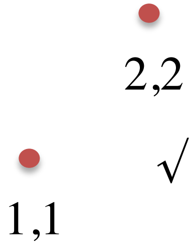
$\sqrt{(2-19)^2}$



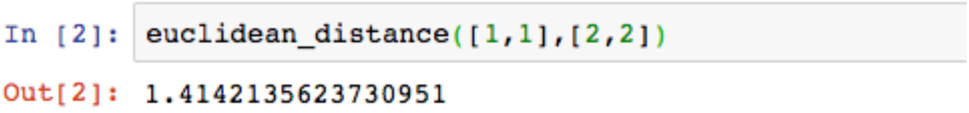
```
In [1]: import math

        def euclidean_distance(input1,input2):
            d = 0
            for i in range(len(input1)):
                d += (input1[i] - input2[i])**2
            return math.sqrt(d)

In [2]: euclidean_distance([2],[19])

Out[2]: 17.0
```

Cultural Analytics

Euclidean Distance

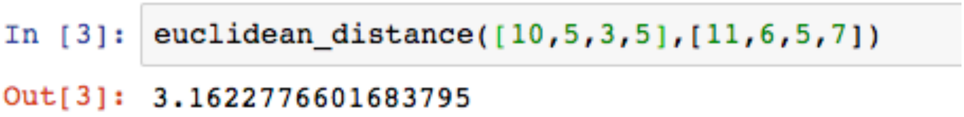


$\sqrt{(1-1)^2 + (2-2)^2}$



```
In [2]: euclidean_distance([1,1],[2,2])
Out[2]: 1.4142135623730951
```

Example from Jockers (Four Features/Words):



```
In [3]: euclidean_distance([10,5,3,5],[11,6,5,7])
Out[3]: 3.1622776601683795
```

Distance Metrics

- **Euclidean**: shortest straight line path
- **Manhattan/City Block**: distance on the grid
- **Cosine Similarity**: measurement of the angle between vectors

Euclidean Distance	Cosine Similarity
0.0 (Frederick Douglass)	-0.0 (Frederick Douglass)
735.075 (Josiah Henson)	0.012 (Josiah Henson)
765.7 (James Watkins)	0.012 (Frederick Douglass)
787.152 (William Wells Brown)	0.012 (Frederick Douglass)
833.014 (Lewis Garrard Clarke)	0.013 (Henry Box Brown)
850.614 (Henry Bleby)	0.013 (Josiah Henson)
873.914 (Olaudah Equiano)	0.014 (Richard Hildreth)
875.796 (Boyrereau Brinch)	0.014 (Richard Hildreth)
883.807 (Richard Hildreth)	0.015 (Frederick Douglass)
914.103 (Okah Tubbee)	0.015 (Austin Steward)
922.094 (Olaudah Equiano)	0.016 (Josiah Henson)
943.941 (James Williams)	0.017 (Lewis Garrard Clarke)
945.584 (Elijah P. Marrs)	0.018 (James W. C. Pennington)
969.875 (James Lindsay Smith)	0.018 (William Wells Brown)
988.361 (Thomas L. Johnson)	0.019 (Henry Bleby)
1039.453 (James W. C. Pennington)	0.019 (Henry Bibb)
1045.759 (Ashton Warner)	0.019 (Lewis Garrard Clarke)
1046.148 (Richard Hildreth)	0.019 (James Watkins)
1055.261 (A. R. Green)	0.02 (John Brown)
1060.764 (John Thompson)	0.021 (James Lindsay Smith)

Jockers, “Influence”

- “My objective now is not to classify novels into nationalities or genders but rather to capture for each book a unique book signal and then to look for signs of historical change from one book to the next” (158).
- “Books are being pulled together (and pushed apart) based on the similarity of their computed stylistic and thematic distances from each other” (164).

Piper, "Fictionality (Sense)"

- **Inquiry question**: What distinguishes fictional from non-fictional prose?
- **Major claims**:
  - "fictionality is a highly legible category at the level of linguistic content…When we take into account a sufficient number of words, we can build predictive models that can identify works of fiction with greater than 95% accuracy" (97).
  - "Given enough words, the intentionality that is supposed to reside beyond the semantic content of a statement is indeed largely recoverable from that semantic content" (97).
  - "It is knowledge, not just of otherness, but of another embodied individual that most consistently frames the epistemological horizon of the novel from a quantitative point of view" (110).

Cultural Analytics

Example Dictionary: Senses

Root -> Sense

```
aroma,smell
fragranc,smell
perfum,smell
pungenc,smell
pungent,smell
reek,smell
scent,smell
smell,smell
sniff,smell
stank,smell
stench,smell
stink,smell
stunk,smell
whiff,smell
```

Total Terms

17	sight
14	smell
18	sound
16	taste
14	touch