Log-odds-ratio

- When comparing corpora, we would like to know whether certain words are more common in one corpus over another while controlling for their relative sizes
- The log-odds-ratio quantifies this notion formally:

$$\text{log odds of } w = \log \frac{\frac{\#\text{ occurences of } w \text{ in A}}{\text{total \# of words in A}}}{\frac{\#\text{ occurences of } w \text{ in B}}{\text{total \# of words in B}}}$$

- To assess behavior, we would like to know whether certain verbs are statistically overrepresented in the event chains for one character compared to the others

Log-odds-ratio: Canonical villains

Draco

```
(('belong',), 1.8800223896830026)
(('marry',), 1.6101027735710618)
(('own',), 1.4500662615665423)
(('knock',), 1.4417674587518468)
(('can',), 1.1975704982398048)
(('hang',), 1.1760642930188414)
(('gaze',), 1.154085386300066)
(('apologize',), 1.1316125304480078)
(('protect',), 1.1316125304480078)
(('shove',), 1.1086230122233083)
(('sink',), 1.0969269724601176)
(('dance',), 1.0363023506436821)
```

Voldemort

```
(('love',), 1.6785005587503603)
(('felt',), 1.3286789422108995)
(('run',), 1.26069807688823552)
(('write',), 1.148417118379955)
(('stand',), 1.1391147257176417)
(('pull',), 1.0647670890930305)
(('head',), 1.0512533699263074)
(('help',), 0.916749687482846)
(('hear',), 0.773723668938544)
(("'m",), 0.744108748399614)
(('smile',), 0.735206038157135)
(('fall',), 0.685793596431842)
```

Directions with log-odds-ratio

- Is vocabulary in fan fiction statistically different from the canon?
- Are certain actions more likely to occur in the fan fiction?
- Are the adjectives used to qualify characters appreciably different across the two corpora?
- Within a story, are certain actions more likely to occur in the earlier chapters than the later ones?