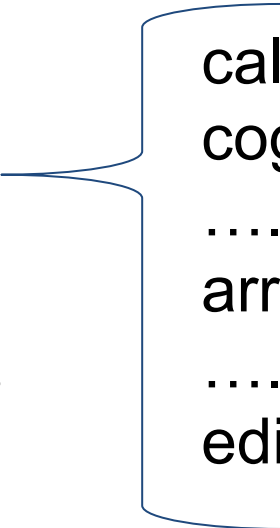


Análise de Mídias Sociais e Mineração de Texto

Avaliação de Modelos

Laura de Oliveira F. Moraes

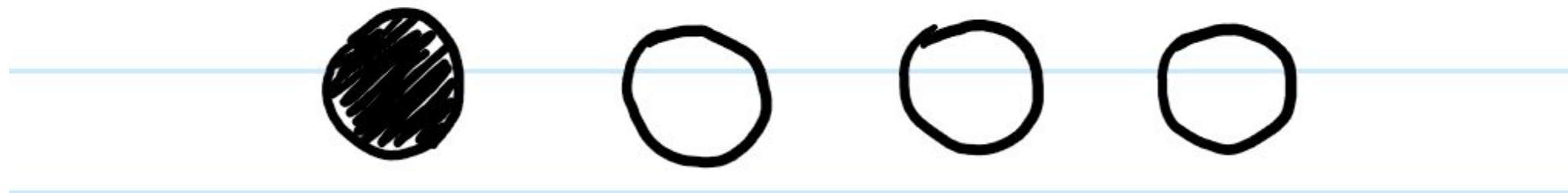
- O quão bem conseguimos prever a próxima palavra?
 - Eu sempre peço pizza com _____
 - O primeiro presidente do Brasil foi _____
 - Eu vi _____
- O melhor modelo de texto é o modelo que atribui a **maior probabilidade** à palavra que na realidade aconteceu.



calabresa 0.1
cogumelos 0.1
.....
arroz 0.01
....
edifício $1e^{-100}$

- É a exponenciação da entropia
- Entropia = **medida de incerteza**

- É a exponenciação da entropia
- Entropia = **medida de incerteza**

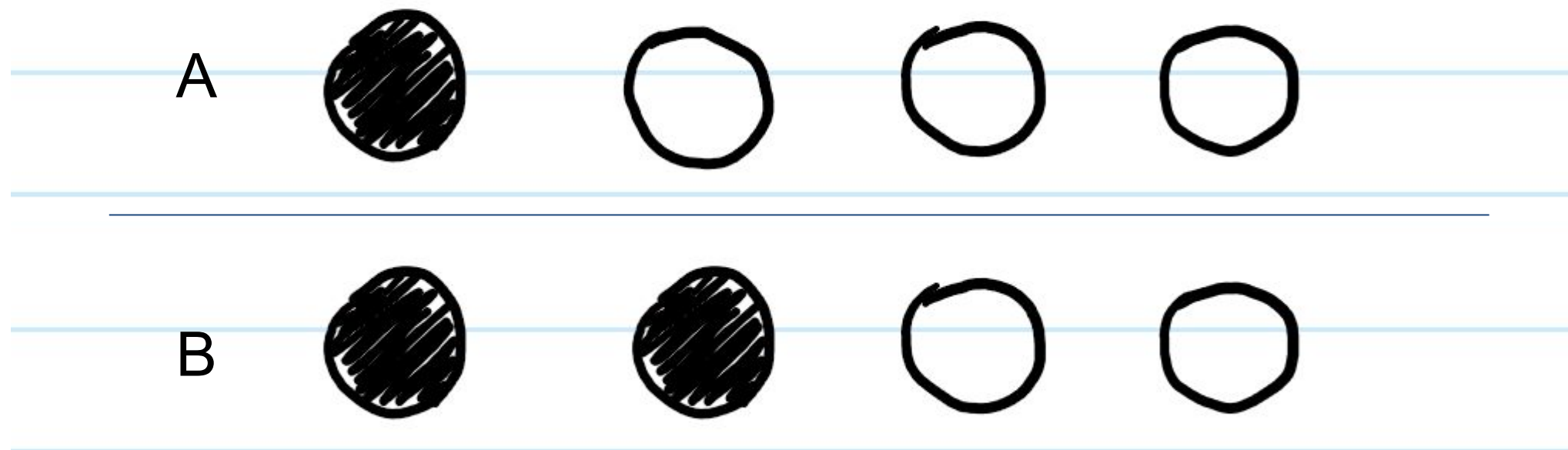


- Qual evento é mais incerto? O preto ou o branco?

- É a exponenciação da entropia
- Entropia = **medida de incerteza**



- Qual evento é mais incerto? O preto ou o branco?



- E a entropia do sistema? Qual você sabe menos o que vai acontecer, a situação A ou B?

- Vimos que nesses modelos, cada tópico pode ser associado a uma lista de **top-N palavras**
- A coerência é a medida da **razão entre a co-ocorrência de palavras no tópico e a ocorrência total**
- A ideia é que se um **tópico é bem-definido**, as mesmas palavras são usadas nos documentos

$$C_{UMass}(t, V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + \epsilon}{D(v_l^{(t)})}$$