1) # of clusters

2) Homogeneity may vary across clusters

3) What features to use?

4) What names to assign to clusters?

   ↳ Dominant?

   ↳ How about leveraging the wiki topic hierarchy for cluster names?

Suggestions: ① Correlate clusters with category ids?

How abt accounting for category structure in clustering process?

Desirable properties of clusters in DAG

① Independence: Cover disjoint doc set/diverse

Loose Orthogonality

② Exhaustive: High coverage

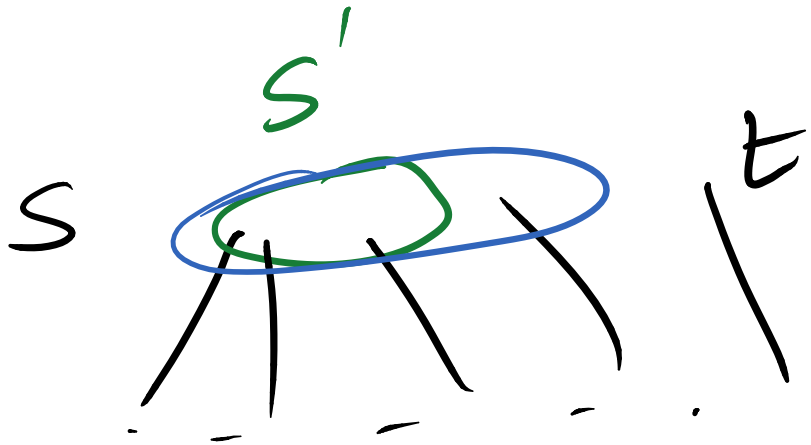
③ Intra cluster similarity · ④ Depth (Quality cnt)

$$\Gamma(S) = \bigcup_{s \in S} \Gamma(s)$$

$$S' \subseteq S$$

$$t \notin S, S'$$

$$|\Gamma(s)| \geq |\Gamma(s')|$$



$\Gamma(s)$, $\Gamma(t)$, $\Gamma(s')$

$\Gamma(s \cup \{t\})$ $\Gamma(s' \cup \{t\})$

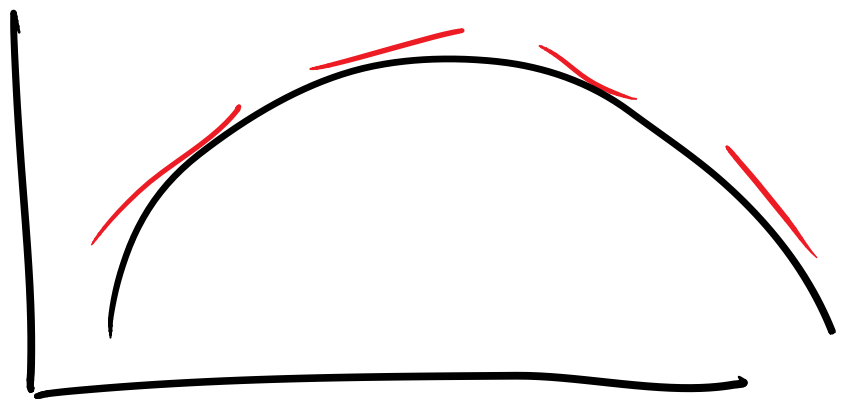
$\mathcal{D}$

$$\Gamma(s' \cup \{t\}) - \Gamma(s') \geq \Gamma(s \cup \{t\}) -$$

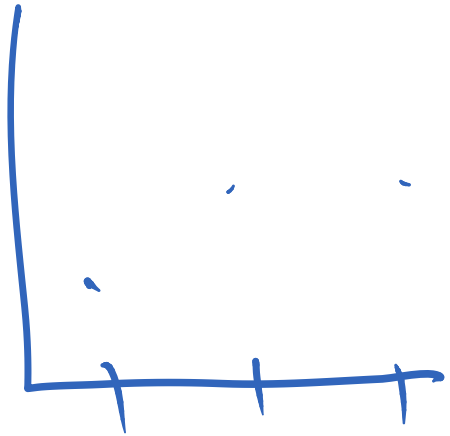
Do submodular fns resemble concave or convex fns in their diminishing returns property?

$S' \subseteq S \subseteq V \setminus \{v\}$

$$f(S' \cup \{v\}) - f(S') \geq f(S \cup \{v\}) - f(S)$$



Slope decreases



$$\Gamma(S) = \bigcup_{s \in S} \Gamma(s)$$

$$\psi(S) = \sum_{s \in S} |\Gamma(s)|$$

Is this submodular?

Submodular

$$m_u(S) = f\left(\bigcup_{s \in S} (s)\right) \rightarrow \sum_{s \in S} m_u(s)$$

$$\psi(S) = \sum_{s \in S} \psi(\{s\})$$

Ans: Modular

$\psi(m_u(S))$ s.t. ψ is concave $\Rightarrow \psi($m