A.

Let $S$ be the total set of pages.

Let $\forall p \in S$: $E(p) = \alpha / |S|$ (for some $0 < \alpha < 1$, e.g. 0.15)

Initialize $\forall p \in S$: $R(p) = 1 / |S|$

Until ranks do not change (much) (convergence)

For each $p \in S$:

$$R'(p) = \frac{1}{q:q \rightarrow p} \sum \frac{R(q)}{N_q} + E(p)$$

$$c = 1 / \sum_{p \in S} R'(p)$$

For each $p \in S$: $R(p) = c R'(p)$ (normalize)

Following the algorithms above, we can get the results for each iteration:

Iteration I:
- Before the normalization: $[A, B, C, D] = [0.0375, 0.0375, 0.5375, 0.2875]$
- After the normalization: $[A, B, C, D] = [0.04166, 0.04166, 0.5972, 0.3194]$

Iteration II:
- Before the normalization: $[A, B, C, D] = [0.0375, 0.0375, 0.120833, 0.6347]$
- After the normalization: $[A, B, C, D] = [0.045152, 0.045152, 0.145489, 0.7642077]$

Iteration III:
- Before the normalization: $[A, B, C, D] = [0.0375, 0.0375, 0.1275, 0.1835]$
- After the normalization: $[A, B, C, D] = [0.097, 0.097, 0.33, 0.475]$

B.

This is an example solution, any other solutions with reasonable explanation are acceptable.

1. Keyword Research is favorable. Keywords are really important when it comes to website ranking. By identifying the keywords the target audience uses, it helps to build effective content for the website to boost the rank.

2. Boosting using term spamming would be frowned on by search engine companies. The presence of terms may result in a match with many queries, increasing the ranking. However, search engine companies would discourage this because it spams user with not useful information.

C.

d1: [Jack London] [London traveled] [traveled to] [to Indianapolis]
d2: [Jack London] [London traveled] [traveled to] [to the] [the city] [city of] [of Indianapolis]
d3: [Jack traveled] [traveled from] [from Indianapolis] [Indianapolis to] [to Chicago]

Jaccard coefficients:

\[ J(d1, d2): \frac{3}{8} \]
\[ J(d2, d3): 0 \]

Pros:
It is a simple model to identify near-duplicates. It also weakly considers context near words by employing the concept of shingles.

Cons:
it is not quite correct when two documents are semantically similar.

D.

When there is no commination among those crawlers, they are very likely to reach same pages and do duplicate work.

E.
Seed the ratings of the new document: (4 pts)
(3 pts for coming up with an idea of how to do so)
a. From ratings of similar documents
b. By showing the document to some selected users and collecting ratings

(3 pts for conveying that seed ratings can be used to suggest the documents to certain users)
Use the seed ratings to find users to show the document to, thus increasing the data available for the document.

(2 pt)
To prevent poor recommendations: ration the number of new documents shown to any particular user

(2 pt)
To make sure the document reaches audience in a timely manner:
a. Get enough seed ratings quickly
b. Include a new documents section for each user

(Total capped to 10 pts)