

CLAIMS

1. A method, comprising:
obtaining, by a processor, a plurality of artifacts in plaintext form;
5 generating, by the processor, fingerprints for each of the plurality of artifacts by
generating shingles from text within each of the plurality of artifacts so
that there at least one character overlap between adjacent shingles; and
cryptographically hashing the shingles to generate a plurality of artifact
fingerprints;
10 storing the artifact fingerprints in a database;
receiving, by the processor from a querying party's system, at least one
fingerprint that has been cryptographically hashed by the querying party's system;
determining, by the processor, whether the received at least one fingerprint
matches any of the artifact fingerprints stored in the database; and
15 outputting, in plaintext form, the artifact containing a matched fingerprint.
2. The method of claim 1, wherein the processor and database are part of a
first system controlled by a first entity, and wherein the first system is a different system
than the querying party's system.
20
3. The method of claim 2, wherein the at least one fingerprint is generated
by:
receiving, by a second processor of the querying party's system, a plaintext
query;
25 generating, by the second processor, a plurality of query shingles from the
plaintext query so that there is no character overlap between adjacent shingles;
cryptographically hashing, by the second processor, the plurality of query
shingles to form a plurality of query fingerprints.

4. The method of claim 3, wherein prior to cryptographically hashing the plurality of query shingles by the second processor, the method further comprises:

comparing, by the second processor, the plurality of query shingles to a stop word list; and

5 removing, by the second processor, any shingle of the plurality of query shingles matching a word in the stop word list.

5. The method of claim 3, wherein prior to determining whether the received at least one fingerprint matches any of the artifact fingerprints stored in the
10 database, the method comprises:

receiving, by the second processor, a set of artificially common fingerprints in the artifact fingerprints; and

removing, by the second processor, any fingerprints from the plurality of query fingerprints matching any of the fingerprints in the set of artificially common
15 fingerprints.

6. The method of claim 3, wherein the shingles generated from the text within each of the plurality of artifacts and the shingles generated from the plaintext query are salted prior to the cryptographic hashing.

20

7. The method of claim 1, wherein the received at least one fingerprint is obtained from the stored artifact fingerprints, the method further comprising:

outputting an indication of an artifact containing the received at least one fingerprint, thereby indicating that the output artifacts contain at least some common
25 text.

8. The method of claim 1, wherein the database storing the artifact fingerprints is a distributed key store.

9. The method of claim 1, wherein the plurality of artifacts are obtained using a web spider that collects websites as it crawls across the internet.

10. A method, comprising:

5 obtaining, by a processor, a plurality of artifacts in plaintext form;
generating, by the processor, fingerprints for each of the plurality of artifacts by
generating shingles from text within each of the plurality of artifacts so
that there at least one character overlap between adjacent shingles; and
cryptographically hashing the shingles to generate a plurality of artifact
10 fingerprints;
storing the artifact fingerprints in a database;
receiving, by the processor, at least one fingerprint;
determining, by the processor, whether the received at least one fingerprint
matches any of the artifact fingerprints stored in the database; and
15 outputting an indication of an artifact containing a matched fingerprint,
wherein the determination of whether the received at least one fingerprint
matches any of the stored artifact fingerprints comprises
generating, by the processor, a plurality of query vectors from the at least
one cryptographically hashed shingle;
20 generating, by the processor, a plurality of artifact vectors from the
artifact fingerprints;
performing, by the processor, a cosine distance calculation using the
plurality of query vectors and the plurality of artifact vectors to generate a
plurality of distance scores; and
25 selecting, by the processor, a vector of the plurality of artifact vectors
having a lowest distance score.

11. The method of claim 10, further comprising:

determining, by the processor, a median cosine distance for the plurality of
30 artifact vectors, wherein the median cosine distance is used to set a threshold against

which the lowest distance score is compared to determine whether the vector of the plurality of artifact vectors having the lowest distance score is a match of one or more of the query vectors.

- 5 12. A method, comprising:
obtaining, by a processor, a plurality of artifacts in plaintext form;
generating, by the processor, fingerprints for each of the plurality of artifacts by
generating shingles from text within each of the plurality of artifacts so
that there at least one character overlap between adjacent shingles; and
10 cryptographically hashing the shingles to generate a plurality of artifact
fingerprints;
storing the artifact fingerprints in a database;
receiving, by the processor, a plurality of query fingerprints, each of which is
generated by cryptographically hashing shingles of a plaintext query, wherein a
15 character overlap between shingles generated from the plaintext query is less than the at
least one character overlap of the shingles generated from the artifacts;
determining, by the processor, whether any of the plurality of query fingerprints
matches any of the artifact fingerprints stored in the database by performing a cosine
distance calculation; and
20 outputting an indication of an artifact containing a matched fingerprint.
13. The method of claim 12, wherein
the cosine distance calculation is performed between each of the artifact
fingerprints stored in the database to produce a median artifact similarity score,
25 the cosine distance calculation is performed between each of the plurality of
query fingerprints and each of the artifact fingerprints to produce individual query
fingerprint scores, and
a match between the any of the query fingerprints and any of the artifact
fingerprints is determined based on both the median artifact similarity score and the
30 individual query fingerprint similarity scores.

14. The method of claim 12, wherein there is no character overlap between shingles generated from the plaintext query.

5 15. The method of claim 12, wherein the processor and database are controlled by a first entity and the at least one fingerprint is received from a second entity, and wherein the plurality of query fingerprints are generated by:

- receiving, by a second processor from the second entity, a plaintext query;
- generating, by the second processor, a plurality of query shingles from the
- 10 plaintext query so that there is no character overlap between adjacent shingles;
- cryptographically hashing, by the second processor, the plurality of query shingles to form the plurality of query fingerprints.

15 16. The method of claim 15, wherein prior to cryptographically hashing the plurality of query shingles by the second processor, the method further comprises:

- comparing, by the second processor, the plurality of query shingles to a stop word list; and
- removing, by the second processor, any shingle of the plurality of query shingles matching a word in the stop word list.

20

17. The method of claim 15, wherein prior to determining whether any of the received query fingerprints match any of the artifact fingerprints stored in the database, the method comprises:

- receiving, by the second processor, a set of artificially common fingerprints in
- 25 the artifact fingerprints; and
- removing, by the second processor, any fingerprints from the plurality of query fingerprints matching any of the fingerprints in the set of artificially common fingerprints.