



Problem Definition

- For natural language applications, we need more training data.
- For language data, the World Wide Web is the ultimate fire hose.
- But how do you drink from it?



Overview

- Our motivation: Chinese text segmentation.
- Our goal: Chinese lexical development.
- Our technique: unsupervised machine learning.
- Our means: corpus collection from the WWW.



Chinese Text Segmentation

- Chinese is written without spaces between words.
- Most natural language applications operate on words.
 - Search engines
 - POS assignment
 - Named entity extraction
 - Parsing
 - ...et cetera.
- So, we need to find the word boundaries.



Segmentation Example

只有为人民工作才有真正价值。



Zhǐ yǒu wèi rén mín gōng zuò cái yǒu zhēn zhèng jià zhí
Only for people work only have real value

Green = correct segments

Red = in lexicon, but not correct here.

There are 80 (eighty!) distinct ways to segment this sentence.
Most of them are nonsense.



Chinese Word Segmentation Approaches

- Purely lexical.
- Purely statistical.
- Hybrid approaches.

They all need a substantial training corpus, annotated by hand.



A 鸡和蛋 Problem

- We need tokenization to get a lexicon and lexical statistics.
- We need a lexicon and lexical statistics for the segmenter.

Which comes first?

In practice, bootstrap with a pre-existing "core" lexicon.



Training With an Annotated Corpus

- We learn (and test) our system based on finding regularities in a corpus.
- If the corpus has been segmented by consistent people, then it can learn what to look for.
- Corpus Limitations:
 - Creation requires huge amount of labor, time and money
 - It goes stale and gets out-of-date
 - Works best only within (limited) domain, degrades outside it
 - ...and there's never enough.



Unsupervised Training for Chinese Word Segmentation

- “Unsupervised”
 - Means “plain text, no human annotation.”
 - Means we don’t know where the words are.
- Look for character sequences.
- If they occur more often than expected, it’s a (possible) word



Problems with Unsupervised Training

- Problem: How to balance likelihoods:
 - Of being a token vs. not.
 - Of being a token in current context.
 - Of the possibility of a new word.
- Problem: finding additional lexical information.
 - POS tags.
 - Simplified vs. traditional versions.
- Problem: find a definition for "word" and consistently apply it.



Other Uses for Unsupervised Training

- Our true goal: look for new words in a document stream.
 - By-product of Unsupervised Word Segmentation.
 - Works continuously.
 - Can be alert for a "burst" signaling a new word.

- Other unsupervised techniques.
 - Clustering words and contexts for classification or POS assignment.
 - "Co-training" for automated finding of names.
 - Clustering documents and assigning topics.



Overview: Building a Large Chinese Corpus

- Crawling
- Post-Processing
 - Strip HTML to get text
 - Remove duplicates
 - Discover document's language and encoding
 - Prepare for further processing



Crawler Desiderata

- Juggling $\gg 10^6$ URIs.
- Minds its manners.
 - Obeys robot-exclusion preferences.
 - Doesn't thrash sites.
- Directed downloading.
 - We're only interested in text - Chinese text, in fact.
 - Everything else is wasted bandwidth and storage.



Crawler Desiderata (2)

- Access to the data while the crawl continues.
 - Since we don't want to wait for the crawl to end.
 - Since the goal is to crawl continuously...
- Recrawling rapidly changing or dynamic sites.
 - News
 - Blogs



Crawling System Software

Crawler: Heritrix 1.4-pre

Language/Encoding ID: Basis Technology's RLI

HTML to text: Vilistextum



Crawl Setup

- Start with 1,500 randomly selected ODP URIs.
- Sun JDK 1.4.2 (512 MB heap)
- 666MHz dual CPU, 1GB RAM
- Started with 50 threads, increased to 150.



Statistics on the First Large Chinese Crawl

| | |
|-----------------------|-----------------------|
| URIs stored: | 7,372,351 |
| URIs stored per sec: | ~8 |
| ARC files: | 300 |
| Total ARC File Size: | 28 GB |
| | |
| Unique Hosts Crawled: | 4032 |
| | |
| Total HTML size: | 109.7 GB |
| Total Stripped size: | 15.8 GB (~5k per doc) |
| | |
| Languages found: | 28 |



Languages Found on First Large Chinese Crawl

| | | | |
|---------------------|-----------|------------|----|
| Simplified Chinese | 5,510,748 | Romanian | 52 |
| Traditional Chinese | 50,030 | Persian | 38 |
| Russian | 5,986 | Hungarian | 32 |
| Japanese | 4,059 | Finnish | 28 |
| Korean | 393 | Bulgarian | 26 |
| Arabic | 365 | Spanish | 11 |
| Polish | 198 | Albanian | 11 |
| Greek | 136 | Vietnamese | 10 |
| Thai | 120 | Swedish | 8 |
| Turkish | 83 | Latvian | 5 |
| Czech | 67 | German | 5 |
| Portuguese | 65 | Icelandic | 3 |
| Hebrew | 58 | Slovak | 2 |
| Lithuanian | 55 | French | 1 |



Finding Unknown Words in Chinese

- The first use for all this data.
- The goal: to find words in Chinese that aren't in our lexicon.
 - Neologisms.
 - New personal names (e.g. celebrities or newsmakers).
 - New foreign words or names borrowed into Chinese.



Word Finding Algorithm

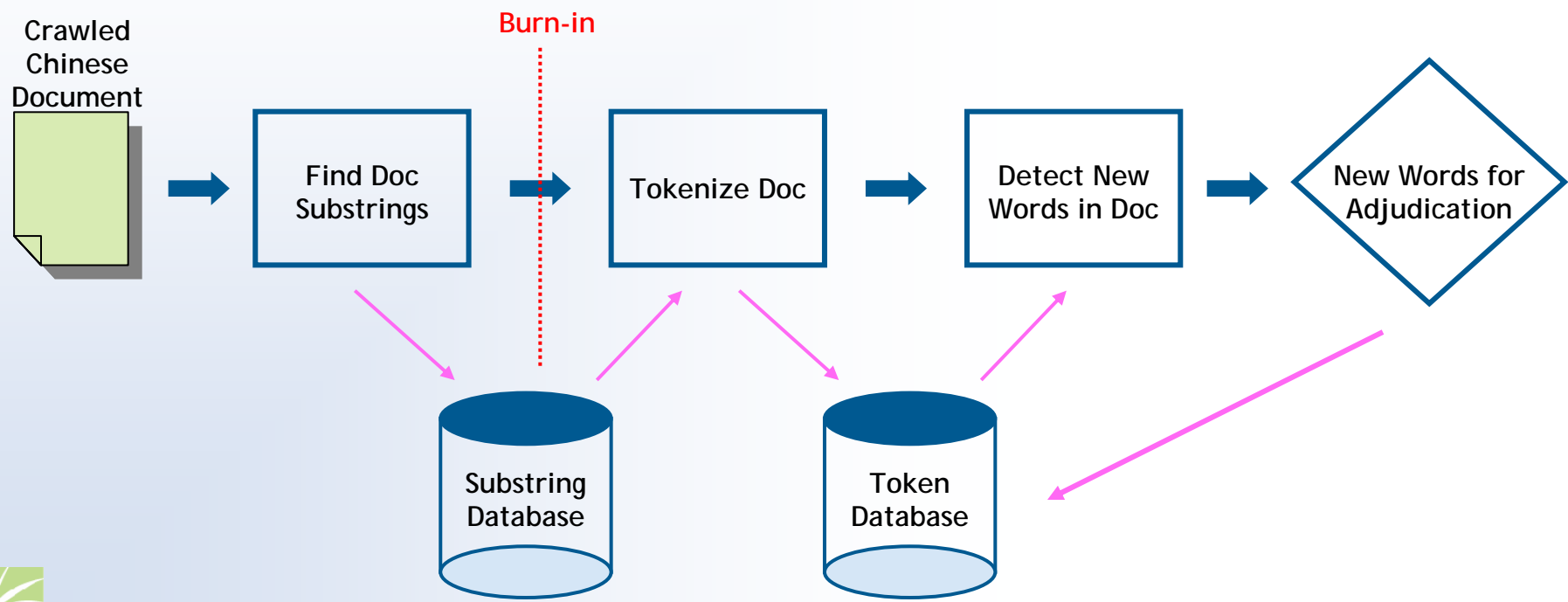
- An on-line algorithm
 - Adapted from the batch algorithm of Jin et.al.
 - Intended to sit on a constant document stream.
- Proceeds incrementally -- with each new document:
 - The internal data is updated.
 - The document is tokenized.
 - One or more new words might be proposed.



Algorithm Overview

For each document in a document stream:

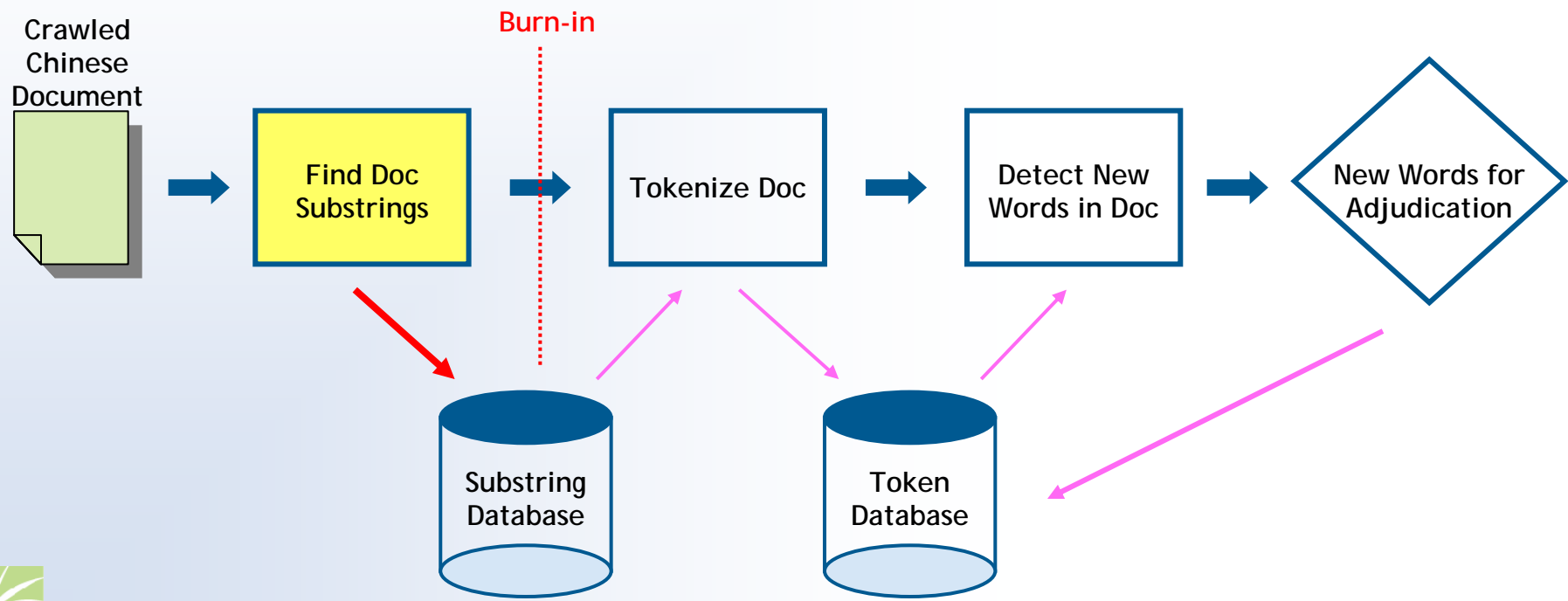
1. Find already-seen substrings within documents.
2. Segment using the substring frequencies.
3. If new words found, propose to human adjudicators.





Step 1: Find Substrings

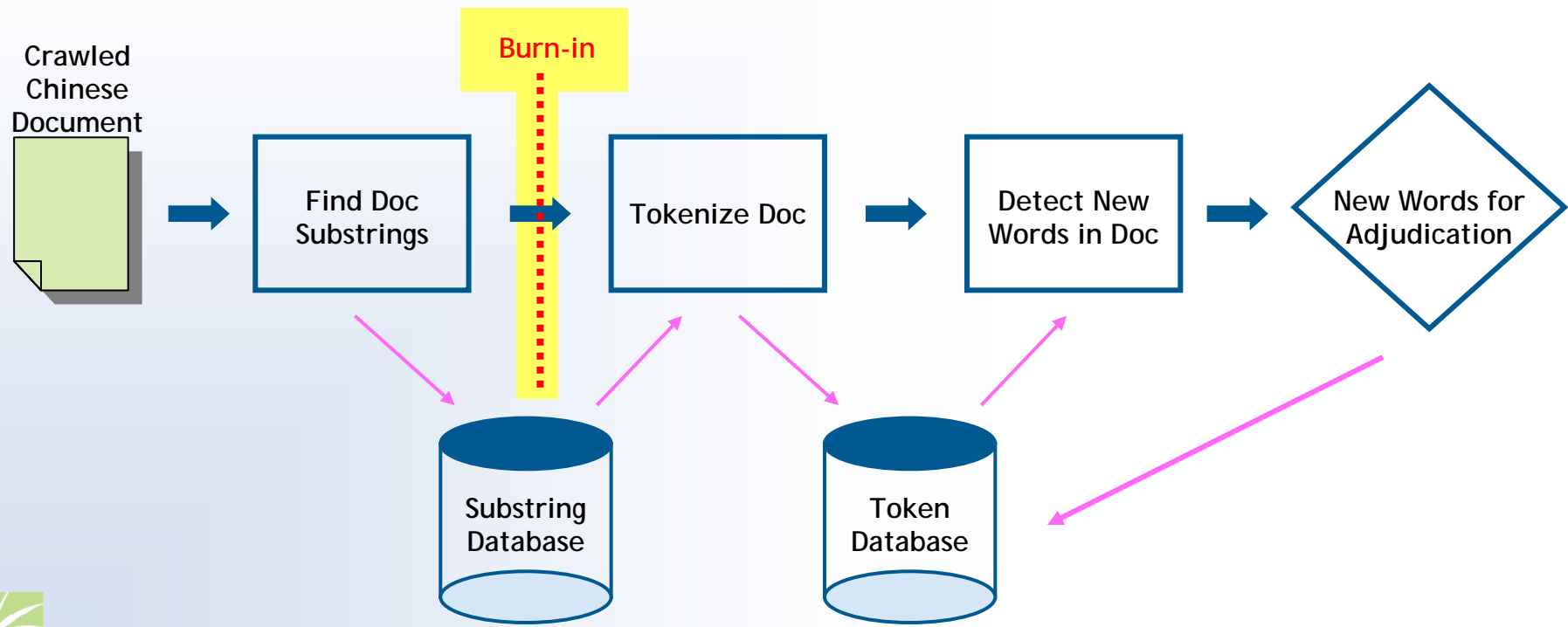
- A map of substring and counts is maintained.
 - Substrings must be of length $2 \leq N \leq 12$.
 - Look for all substrings in a document.
 - If it occurs in two or more sentences, it's added to the substring DB.





Burn-In

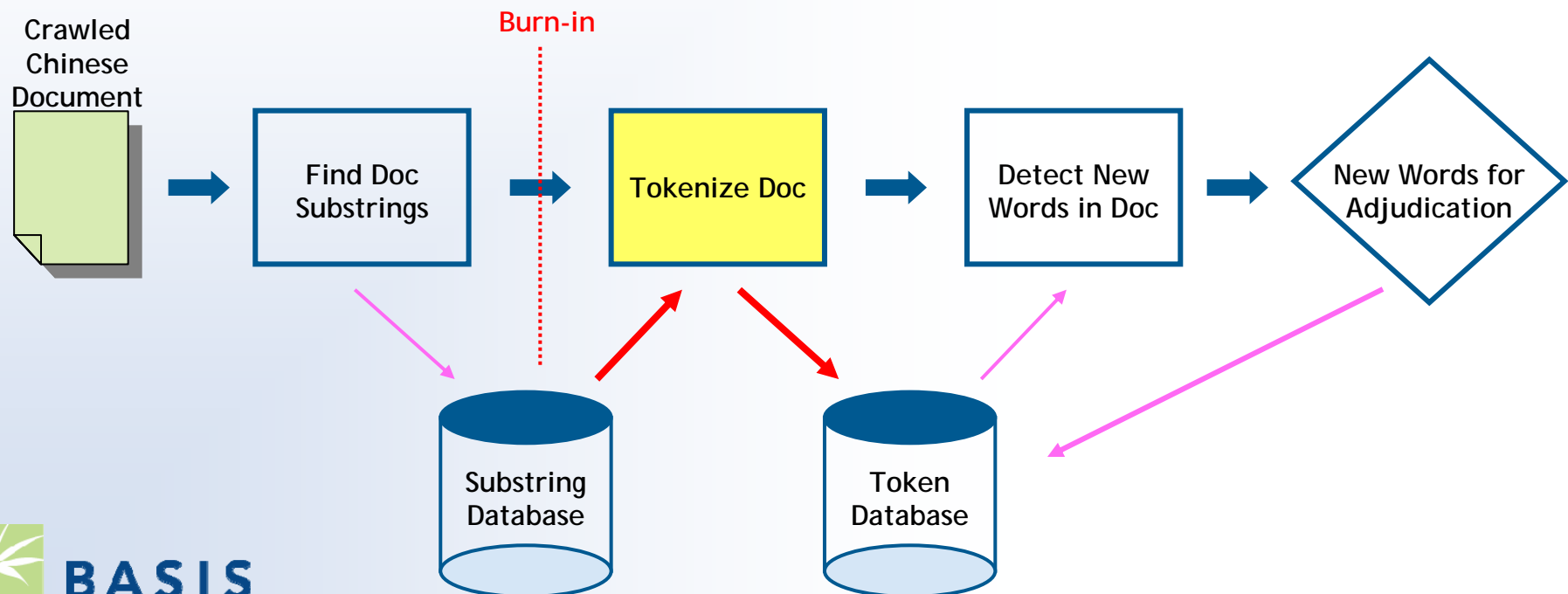
- During burn-in, only substrings are collected.
 - Burn-in length is set by the user.
 - At least tens of thousands of documents.





Step 2: Use substrings to segment into words

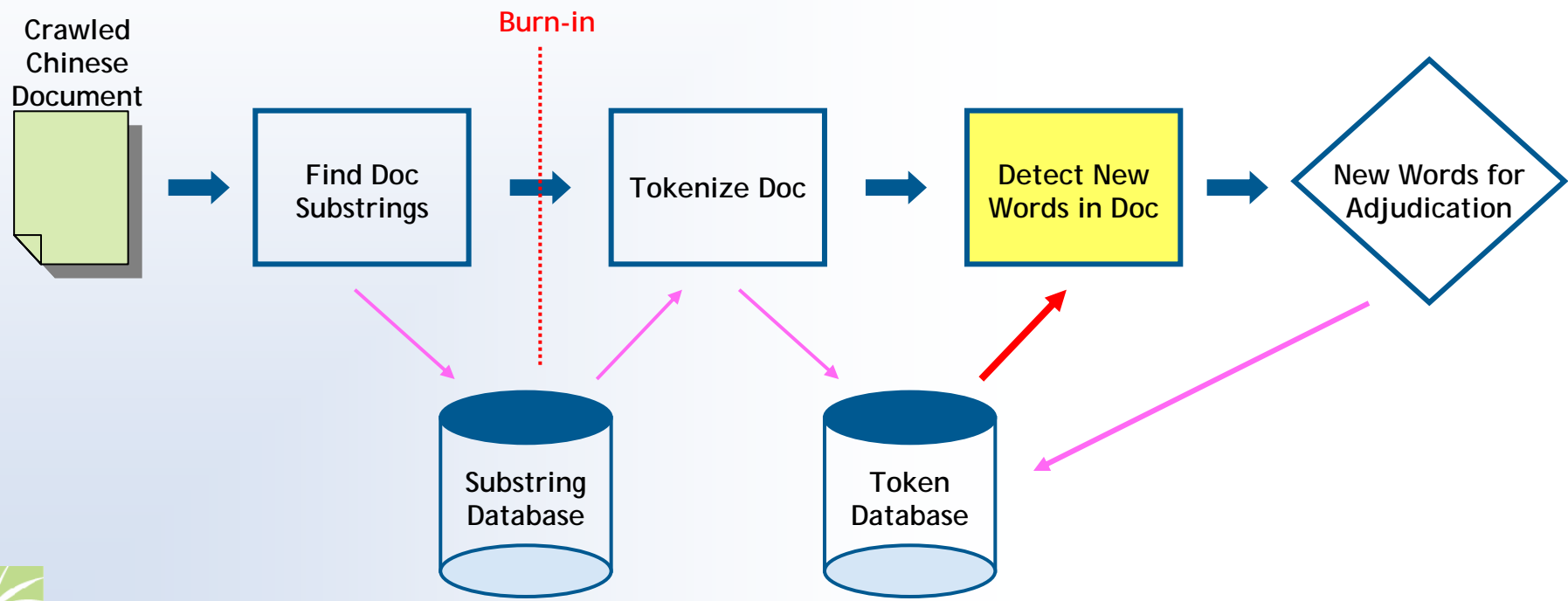
1. In each sentence, find all the substrings and counts.
2. Remove substrings, starting with the lowest-count ones.
3. Leave a substring if removing it would make a >1 character gap.
4. Repeat (2)-(4) until no substrings overlap.
5. For each token found, add one to its count in the Token DB.





Step 3: Detect new words

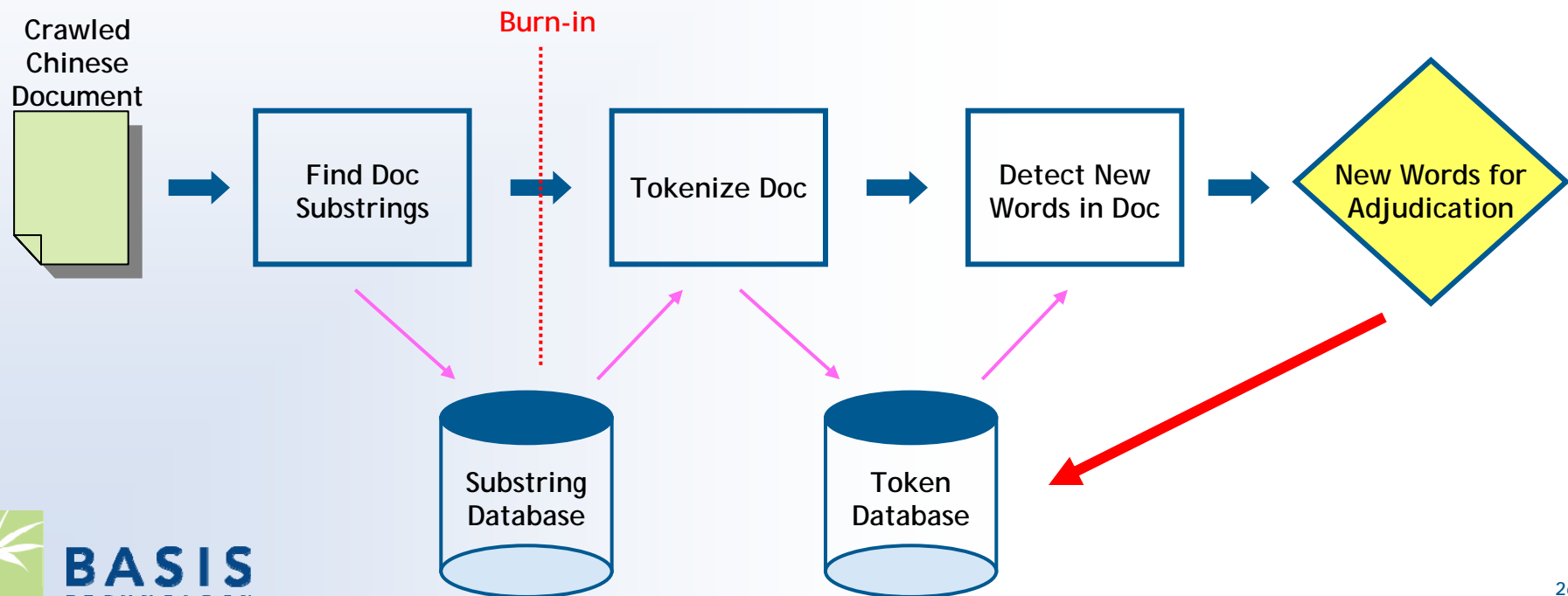
- Check updated token database for possible new words
 - Not already in lexicon.
 - Has occurred more than N times; threshold varies with # docs
 - Favor more recent occurrences to recognize “bursts”

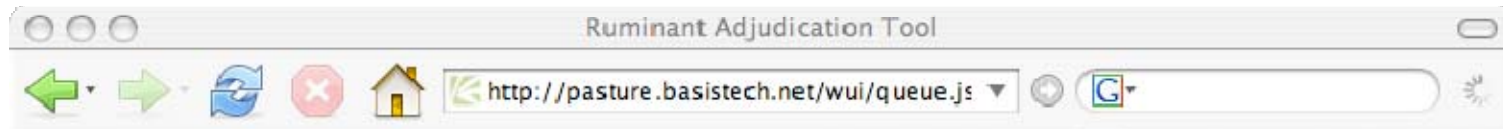




Step 4: Adjudicate new words

- Native Chinese speakers decide if the proposed words are "real".
- They also add other information:
 - Part of speech.
 - If it's a person's name, organization name, or location name.
- Feedback adjudicator's decision to system databases.





Ruminant Adjudication Tool

Dashboard | Queue | Completed | Logged in as tree

Adjudicate a lexeme from the queue [\[refresh\]](#)

| | | | |
|-------------|--------------|----|------------|
| <h1>等也</h1> | 政兼內政部政務部長何炳基 | 等也 | 將隨行 |
| | 李必賢 | 等也 | 紛紛表示支持他的主張 |
| | 尾燈和營業小客車頂燈 | 等也 | 都以條文明文規定 |
| | 陳光復 | 等也 | 先後登上宣傳車高呼 |
| | 民進黨立委洪奇昌 | 等也 | 陸續來到國安局 |
| | 民進黨立委洪奇昌 | 等也 | 陸續來到國安局 |

Please adjudicate this lexeme:

| | |
|-----------------------|--|
| Validity | <input checked="" type="radio"/> VALID <input type="radio"/> INVALID |
| Parts of Speech | <input type="checkbox"/> Abbreviation <input type="checkbox"/> Adjective <input type="checkbox"/> Adverb <input type="checkbox"/> Common noun <input type="checkbox"/> Construction <input type="checkbox"/> Direction word <input type="checkbox"/> Generic noun <input type="checkbox"/> Noun phrase <input type="checkbox"/> Numeric <input type="checkbox"/> Onomatope <input type="checkbox"/> Other <input type="checkbox"/> Phrase <input type="checkbox"/> Profanity <input type="checkbox"/> Pronoun <input type="checkbox"/> Proper noun <input type="checkbox"/> Temporal noun <input type="checkbox"/> Verb |
| Decomposition Pattern | 2 |
| Reading | |
| Add a comment | |





Word Discovery Results

- About 40% of proposed words are accepted by adjudicators.
- Several thousand so far.
- Most new words are proper nouns or common nouns.
- Sources of new words:
 - Lexical innovations.
 - Proper names.
 - Traditional Chinese characters mixed into Simplified texts.



Conclusion

- How we compiled a large corpus of Simplified Chinese.
- We also collected a similar-sized corpus of Traditional Chinese.
- Tool chain has worked well for us
- To do:
 - Moving to fully incremental crawling, rather than periodic (large) crawling jobs
 - New languages and functions



The End

Thanks!

If you'd like to ask a question, here are a few to ask:

- Why is it so hard to decide what a word in Chinese is?
- Tell me about your catastrophic data loss.
- How difficult is it to strip HTML, really?