

テキストマイニングソフトを利用した 新未来洞察手法の研究

2013年11月30日(土)
明治大学駿河台キャンパス
リバーティタワー1001教室

目 次

1. 研究概要
2. 第10分科会メンバー
3. 2011年度、2012年の研究内容と課題
 - 1) 2011年度の研究内容と課題
 - 2) 2012年度の研究内容と課題
4. 2013年度の研究プロセス(未来洞察ツール開発)
5. 3種の未来予測手法結果の比較
6. 2013年度の研究内容のまとめ
7. 参考
 - 1) デルファイ法による菓子のトレンド予測 チーム-1
 - 2) 統計手法による菓子のトレンド予測(参考)

1.研究概要

1.現状、経緯

- 現在、日本の消費財メーカーにおいては、過去からの延長線で描いた未来に向けた施策では立ち行かないとの懸念から、未来洞察(フォーサイト)研究が求められている。
- しかし、未来洞察手法は、デルファイ法、シナリオ法等第二次世界大戦時に米国で開発された専門家による予測手法に頼っている。

2.あるべき姿

- イノベーションが頻繁に起こる科学技術分野においては、現状の専門家による予測手法がツールとして最適と思われるが、食品等の消費財分野では、消費者ニーズを基点としたマーケティングが定着する現在、消費者ニーズを基にした未来洞察手法開発が求められている。

3.ギャップ、ソリューション

- ネット上には、膨大な数の消費者の発言・声がある。これらの中から、当該市場の未来に関わる声をどの様なソフト、手順で集め、集めた声をテキストマイニングソフトを使用し、どのように未来予測として表すか。
- また、これらを個人のネット環境で、フリーソフトだけの利用で実現できるか。

4.研究テーマ(5年後の菓子トレンド)

- 「消費財分野の未来洞察ツール開発」をテーマとし、
 - 1) 専門家による従来の未来洞察手法(デルファイ法)
 - 2) 統計手法による予測(食マップ)
 - 3) ネット上にて、一般消費者が書き込んでいる当該テーマ(菓子)に関するニーズ・不満を収集し、テキストマイニングソフトにて分析。(新未来洞察ツール)
- 1)、2)、3)を比較し、確からしさを考察・分析することで、一般消費者のニーズ発言を基にした新未来洞察ツールの可能性を考察する。
- 2013年は、さらに「新未来洞察ツール」が、個人のネット環境下でフリーソフトだけの使用で可能を考察する。

2.第10分科会メンバー

チーム 1

■東急エージェンシー	志智 氏
■(株)ライオン	新條 氏
■ハウス食品(株)	宮戸 氏
■オフィス東井	東井 氏
■(株)亀田製菓	嶋 氏
■もんじゅ	江澤 氏
■(株)ロッテ	須藤

- 食関連の専門家チーム
- 従来のデルファイ法により、「5年後の菓子トレンド」を予測

チーム 2

■縄文コミュニケーション	福田 氏
■(株)東芝	後藤 氏
■(株)コプロシステム	大久保 氏
■	空閑 氏
■トランスコスモス	北出 氏
■富士通	後藤 氏 安藤 氏
	田中 氏
■R&D	五條 氏
■ロッテ	荒生

- リサーチ関連の専門家チーム
- フリーソフトでのブログからのテキストデータの収集とフリーのテキストマイニングソフトを使い、「5年後の菓子トレンド」を予測

3.2011年度、2012年度の研究内容と課題



1).2011年度研究内容と課題

テキストデータ収集

- ネット上に「菓子未来研究所」を設立し、消費者ニーズを中心に未来のお菓子に係わる発言を収集。

テキストマイニング

- 見える化エンジンを用いて、最頻出ワード順に「5年後の菓子10大トレンド」として編集。

結果

- 直接消費者から収集したテキストデータによる「5年後の菓子10大トレンドは、従来の「デルファイ法」、「統計データ予測」と比較して、それぞれ近似の予測が5トレンドずつ、全く同じ内容も2トレンドずつと遜色ない内容であった。

⇒ 一般消費者の声による未来洞察は専門家の未来洞察と同等

課題

- 直接消費者からの収集ではなく、ブログ等の発言は未来洞察のテキストデータとして有効か？
- どの様に、ブログ等から未来洞察に有効なテキストデータを収集するか？

2).2012年度研究内容と課題

テキストデータ収集

- (株)富士通の協力の基、2. 5億ブログ記事の中から、約1000の未来洞察に有用と思われるテキストデータを抽出

テキストマイニング

- ベクストマイナー(電通盤)用いて、最頻出ワード順に「5年後の菓子10大トレンド」
として編集。

結 果

- ブログから収集したテキストデータによる「5年後の菓子10大トレンドは、従来の「デルファイ法」、
「統計データ予測」と比較して、それぞれ近似の予測が5トレンドずつ、全く同じ内容も2トレンドずつと
遜色ない内容であった。

⇒ ブログの中にも、未来洞察に有用なテキストデータがある。

課 題

- 個人のサーバー環境では、ブログから有用なテキストデータを抽出できない。
- 個人のサーバー環境化でブログ等から未来洞察に有効なテキストデータを収集するフリーツールの研究。

4、2013年度研究プロセス



2013年研究プロセス

【Step1】

ブログ人気ランキングを参考に40の個人ブログの選定

【Step2】

フリークローリングツール「巡集」でブログ記事の収集

【Step3】

収集したブログ記事を「TextBinder」、「TextSS. Net」で1行ずつに加工

【Step4】

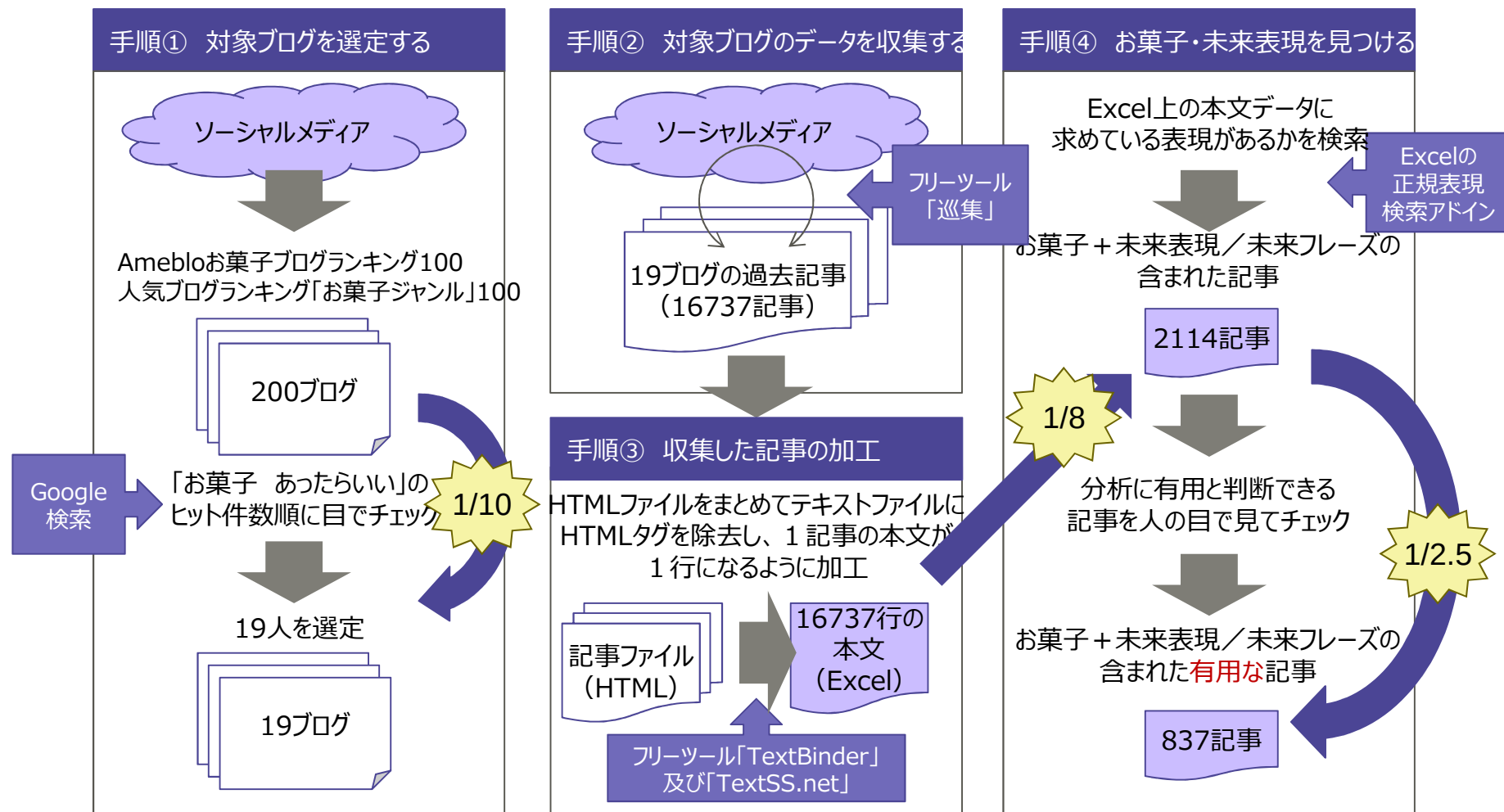
Excel検索機能を用いて、有用なテキストデータの検索

【Step5】

収集したテキストデータをフリーテキストマイニングソフト「KHCoder」で編集

【STEP.1】 ブログでの投稿データ収集 — 記事収集の手順

■ ブログから、テキストマイニングに用いる投稿データを収集する手順を示します。



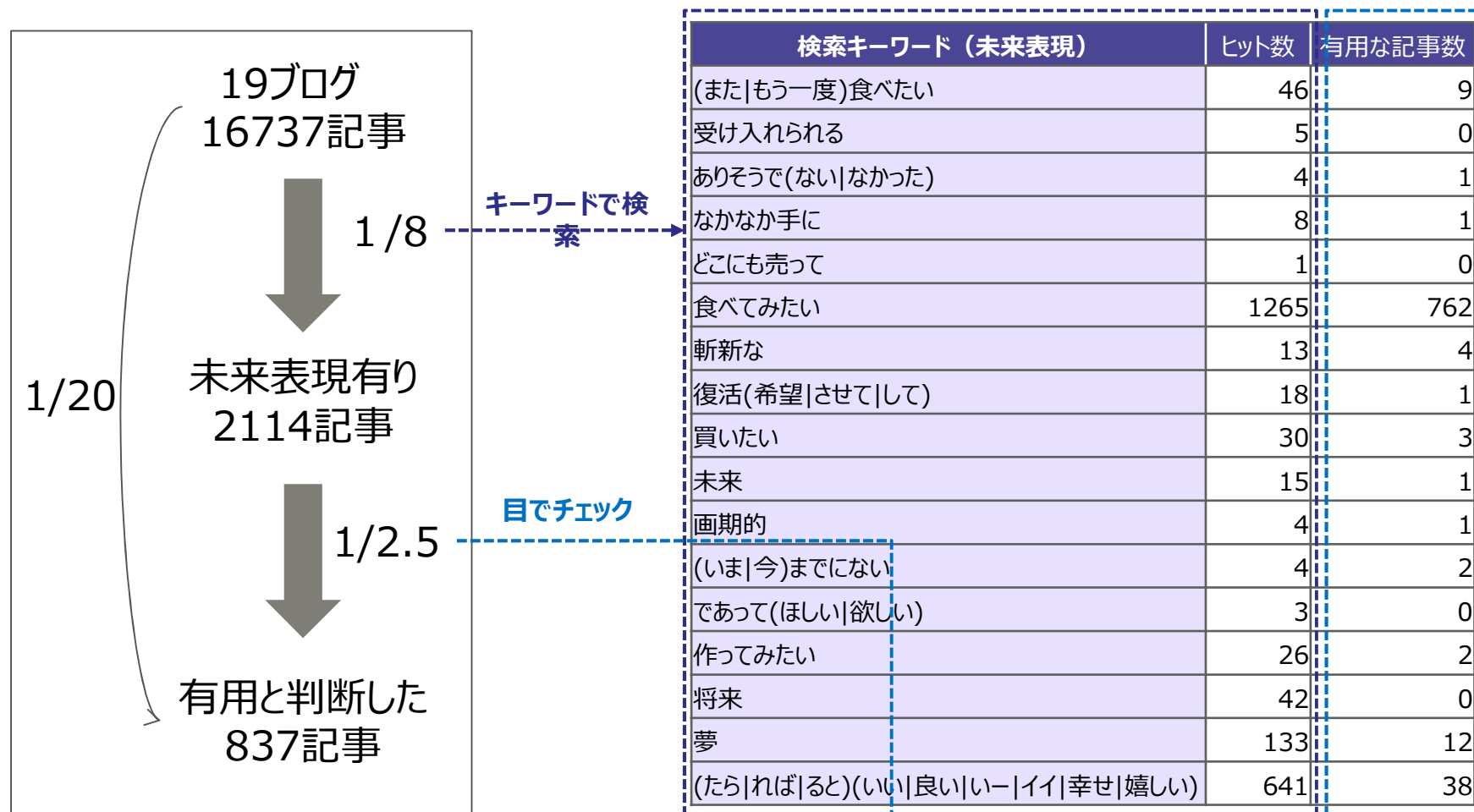
【STEP.1】 ブログでの投稿データ収集 ― 記事収集の対象ブログ一覧

■ 記事収集の対象とした19ブログの一覧です。

No.	ブログ	ブログ名	U R L	記事数	
1	Ameblo	* スイート ローズ * 幸せのひととき♪	http://ameblo.jp/morning-sweet/	260	5168
2		仁美お菓子教室のブログ	http://ameblo.jp/ecole-hitomi/	289	
3		柚猫日記	http://ameblo.jp/yuzuyu220/	270	
4		コンシェル入ってる？	http://ameblo.jp/unirize/	70	
5		ココロとカラダを元気にするお菓子を求めて！	http://ameblo.jp/grainsdenature/	326	
6		マミーのブログ	http://ameblo.jp/ayumimommy/	1204	
7		バカなカエルは木に登る	http://ameblo.jp/natsukyabetsu/	51	
8		chibinaoはスイーツがめっちゃ好きやねん♪	http://ameblo.jp/chibinao0918/	1086	
9		お菓子作家HAPPYの子育て日記	http://ameblo.jp/deephappiness/	927	
10		あまおうたん♪のスイーツ&グルメブログ	http://ameblo.jp/amaku-setsunai/	272	
11		追い続ければきっと願いはかなう。	http://ameblo.jp/hashiko845sskb/	269	
12		まりのブログ	http://ameblo.jp/sora818angel/	144	
13	seesaa	気になる新商品	http://kininarushinshouhin.seesaa.net/	1326	3538
14		*** お菓子なアメリカ***	http://americansweets.seesaa.net/	823	
15		ハタラク蟻さんの生活日記	http://w-ari.seesaa.net/	1314	
16		シフォンケーキ王国！	http://chiffoncake-maple.seesaa.net/	75	
17	exblog	R-Sweetsな生活	http://Rsweets.exblog.jp/	1186	5606
18		petit cafe	http://petitp.exblog.jp/	3101	
19		どすこいスイーツ道場	http://miwadorami.exblog.jp/	1319	
			合計	16737	

【STEP.1】 ブログでの投稿データ収集 — 未来表現の書かれたブログの絞り込み

- 未来表現の含まれる記事の絞り込みに用いた検索キーワードについて、キーワードの含まれる記事の数(ヒット数)と、その中で有用であった記事の数(有用な記事数)を整理しました。



必要なツールと入手先

■ 手順①～④について必要なツールと入手先を記載します。

手順	用途	ツールと入手先
手順① 対象ブログを選定する	ジャンル・カテゴリ別に人気ブロガーを見つける。	Amebloのジャンルランキング一覧： http://ranking.ameba.jp/genre/ 人気ブログランキング： http://blog.with2.net/ その他のブログサイトでもジャンルランキングは存在する。
	「お菓子＋未来表現」が書かれているかチェックする。	Google検索（オプション機能を利用）： http://www.google.co.jp/
手順② 対象ブログデータ収集	指定したURLのブログ記事ファイルをダウンロードする。	巡集： http://www.vector.co.jp/soft/win95/net/se314781.html
手順③ 収集した記事の加工	HTMLファイルをまとめて1つのテキストファイルにする。	Text-File-Binder： http://www.asahi-net.or.jp/~gj3t-tki/REALbasicAPL/TextFileBinder.html Windows8の場合： http://www.vector.co.jp/soft/winnt/util/se377447.html
	HTMLタグを除去したり、1記事の本文が1行になるように加工（置換）する。	TextSS.net： http://www.vector.co.jp/soft/dl/winnt/util/se493374.html
手順④ お菓子・未来表現検索	Excel上で正規表現※でキーワード検索を行う。	Excel用正規表現検索ダイアログアドイン http://srcedit.pekori.jp/tool/excelre.html

※正規表現とは、文字列をパターンとして表現する方法で、これを駆使すると複雑な条件で検索、置換を行うことができます。

手順① 対象ブログを選定する — テーマ(ブログのジャンル)の選定

- 対象ブログを見つけるには、次の方法が有効です。
- Ameblo 人気ブログランキング[ジャンル別](<http://ranking.ameba.jp/genre/category>)、FC2 Blog ジャンル別検索(<http://blog.fc2.com/>)等から、調査対象となるブログテーマを選択してランキングを表示します。
- ジャンルは、各ブログサービスであらかじめ設定されているものの中から、執筆者であるユーザーが任意に選択しています。
- そのため、調査者が求めるジャンルを的確に指定して検索を行うことは、非常に困難です。

◆Amebloの場合

【ジャンルをクリック】



【ランキングをチェック】



◆FC2ブログの場合

【ジャンルをクリックしてランキングを表示】



手順① 対象ブログを選定する — ブログの選定

- 対象ブログを見つけるには、次の方法が有効です。
- 表示されたランキング中のブログを目検でチェックし、目的とする“テーマ専門ブログ”であるか否かを判断します。
- ランキングに表示されるのは、ブログタイトルのみです。ブログタイトルのみから記事の内容を判断することは極めて困難なので、目ぼしいタイトルをクリックし、個々の記事内容を確認する必要があります。
- 選択したブログが、目的とするテーマのみを取り扱った“専門ブログ”であるか否かを判断するためには、複数の記事に目を通す必要があります。

【ブログタイトルをクリック】



【記事内容を目検でチェック】



【ご参考】特定のキーワード・表現のあるブログをあらかじめ絞り込む(1/2)

- 『お菓子』に関する『未来表現のある』ブログなど、明確に収集したいブログが定まっている場合は、Googleの「検索オプション」機能を使用することで、ランキングなどを經由せずに、収集対象となりそうなブログを絞り込むことができます。
- 「検索オプション」機能は、例えば、【「お菓子」を必ず含み、「あったらいい」「将来」「作ってみたい」のいずれかの『未来表現』を含む、Amebloの記事】といった、複雑な検索条件を指定し、それに合致するページを検索する機能です。
- これにより、『お菓子』に関する『未来表現のある』記事を検索することが出来るので、その記事を含むブログを検索対象に定めることができます。
- 「検索オプション」機能は、Googleトップ画面の右上から選択することができます。

【検索オプション機能を呼び出す】



【検索オプション機能選択後の画面】



【ご参考】特定のキーワード・表現のあるブログをあらかじめ絞り込む(2/2)

- 検索オプション機能で、任意の検索条件を設定し、対象とするブログを絞り込みます。
- ここでは、以下の条件を例に手順を示します。「お菓子」を必ず含み、「あったらいい」「将来」のいずれかを含む Amebloの記事
- 「すべてのキーワードを含む」欄に、“必須のキーワード”=「お菓子」を入力します。これにより、「お菓子」を必ず含むWebサイトが検索対象となります。
- 「いずれかのキーワードを含む」欄に、“抽出したい未来表現”である「あったらいい」「将来」を入力します。これにより、「あったらいい」もしくは「将来」を含むWebサイトが検索対象となります。
- 「サイトまたはドメイン」欄に、“対象とするブログ”のURL=「http://ameblo.jp/」を入力します。これにより、Ameblo上のブログ記事が検索対象となります。
- 以上の条件を設定後、「詳細検索」ボタンをクリックすると、すべての条件に当てはまるページが検索されます。検索されたブログの複数記事を目見でチェックし、調査対象として適切な“専門ブログ”かどうかを判断してください。

The screenshot shows the Google Advanced Search interface in Japanese. The browser tabs include 'Google 検索オプション' and 'お菓子 あったらいい OR'. The address bar shows 'www.google.co.jp/advanced_search?hl=ja'. The search options are set as follows:

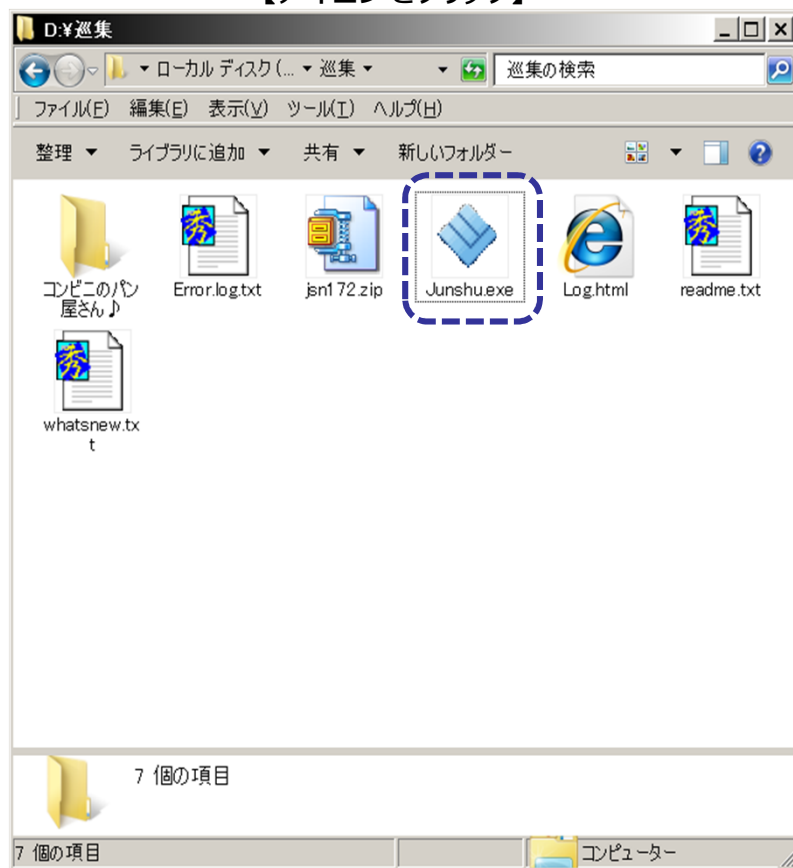
- 検索するキーワード:** 'すべてのキーワードを含む:' field contains 'お菓子'. A callout bubble says: '必須のキーワードを入力 ※ここでは“お菓子”'.
- いずれかのキーワードを含む:** field contains 'あったらいい 将来'. A callout bubble says: '候補となるキーワードを入力 ※ここでは“未来表現”として「あったらいい」「将来」'.
- 検索結果の絞り込み:**
 - 言語:** 'すべての言語'
 - 地域:** 'すべての地域'
 - 最終更新:** '指定なし'
 - サイトまたはドメイン:** 'http://ameblo.jp/'. A callout bubble says: '検索対象とするブログのドメインを指定 ※ここでは、http://ameblo.jp/'.
 - 検索対象の範囲:** 'ページ全体'
 - セーフサーチ:** 'オフ' (with '中' as an option)
 - ファイル形式:** 'すべての形式'
 - ライセンス:** 'ライセンスでフィルタリングしない'

The '詳細検索' (Detailed Search) button is at the bottom right.

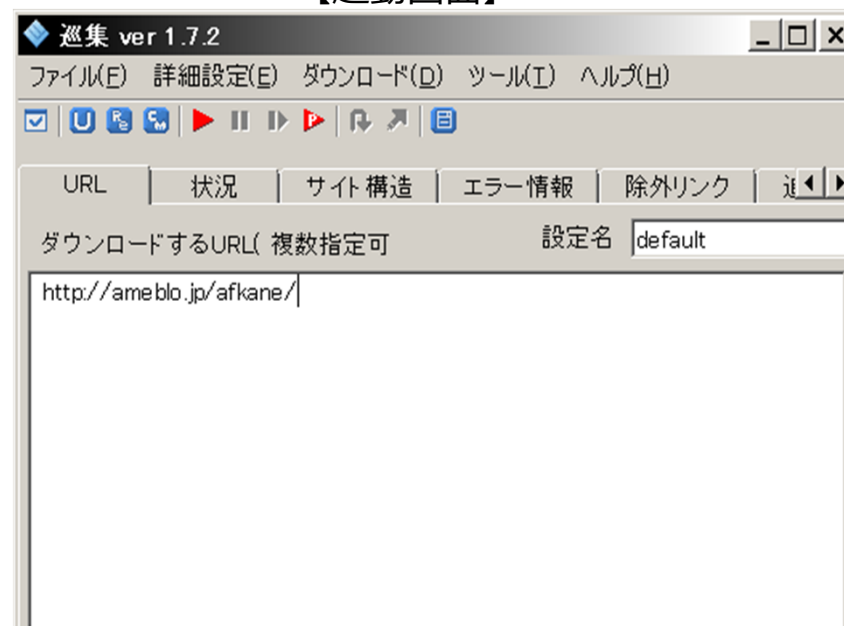
手順② 対象ブログデータ収集 — ツールの起動

- 収集したいサイトが特定できたら、フリーツール「巡集」を起動します。(Junshu.exeをクリック)
- 起動画面は右のようになります。

【アイコンをクリック】

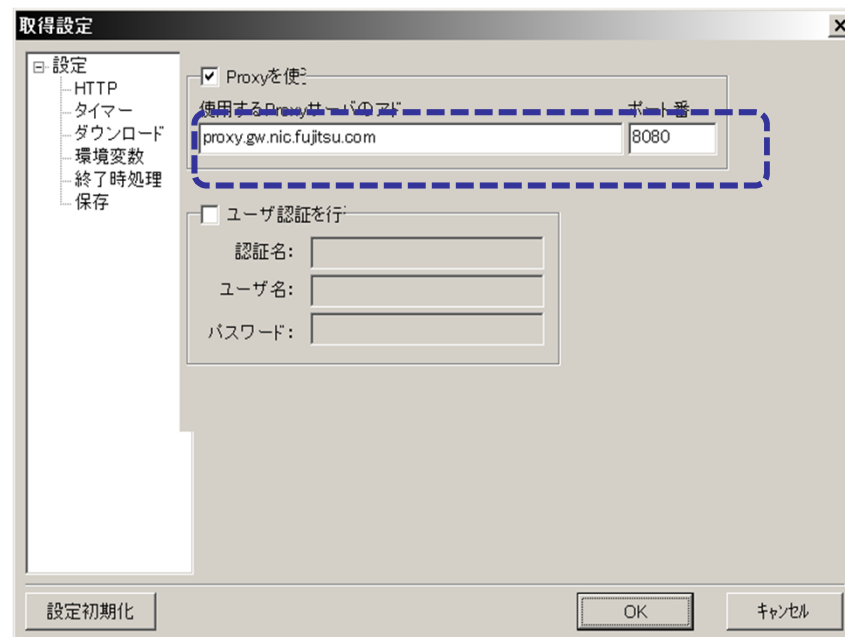
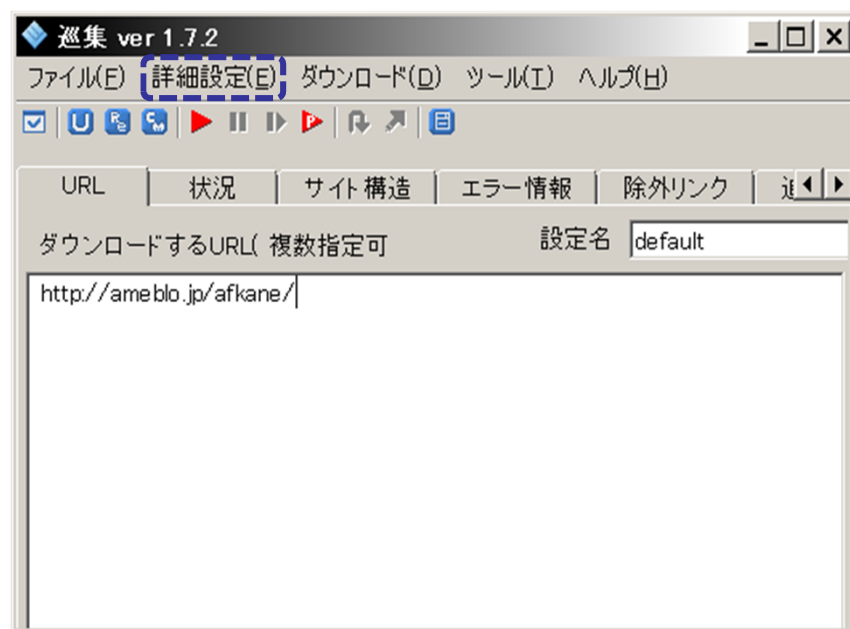


【起動画面】



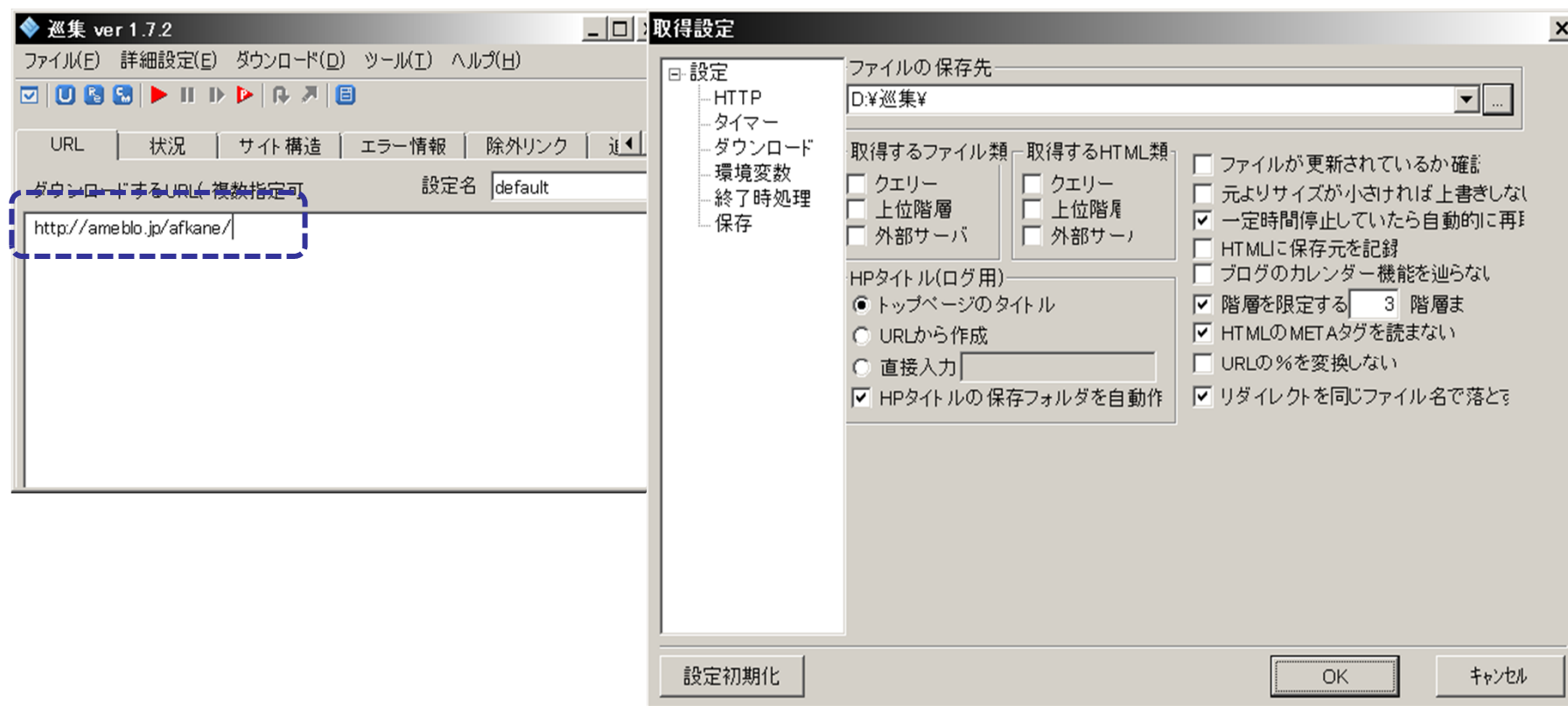
手順② 対象ブログデータ収集 — Proxyの設定

- ツールがネットにアクセスできるように、詳細設定＞取得設定＞HTTPを指定してProxyサーバとポート番号を入力します。(わからない場合はネットワークの管理者に確認。)



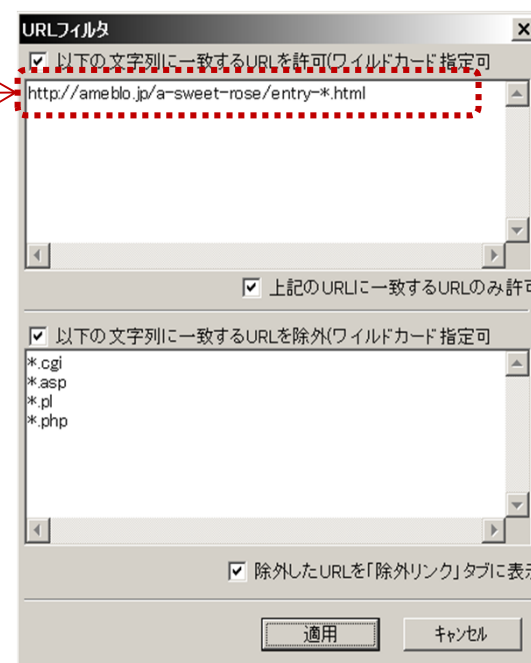
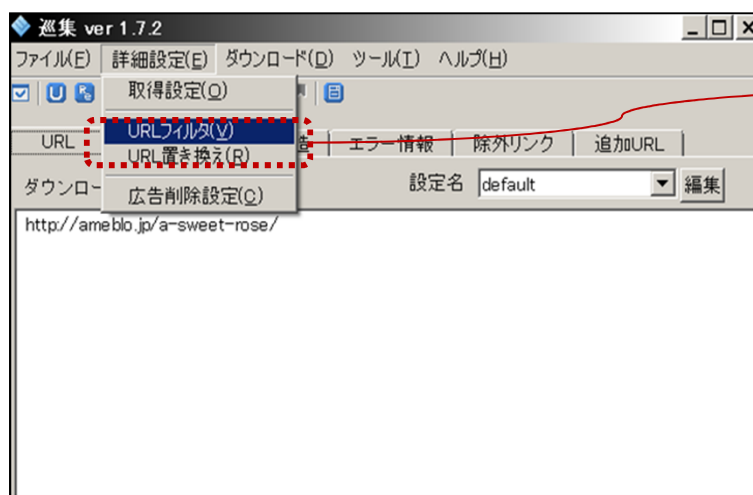
手順② 対象ブログデータ収集 — ダウンロード設定

- ダウンロードしたいブログのトップサイトのURLを、「URL」タブの入力欄に記入します。
- 「詳細設定＞取得設定」でオプションを設定します。
設定の変更が必要な項目は以下の通りです。
 - ファイルの保存先・・・ダウンロードしたファイルの保存先を任意に指定してください。
 - 階層を限定する・・・この項目にチェックを入れ、「1階層まで」に設定します。



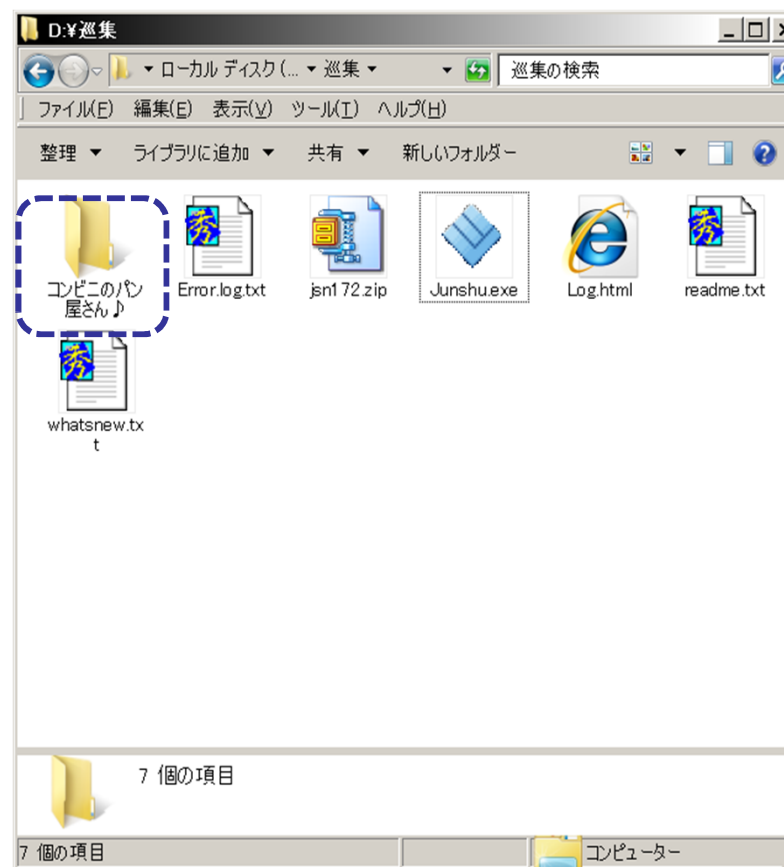
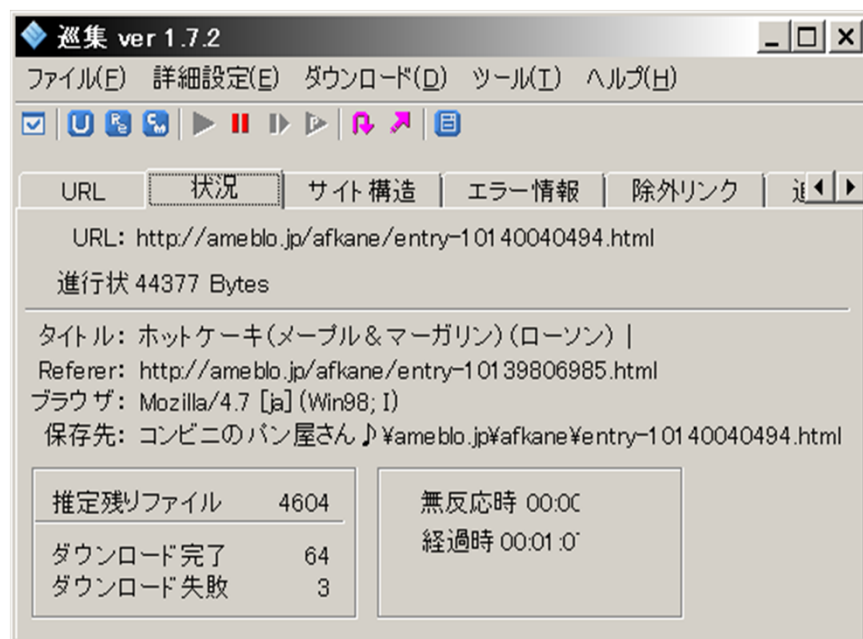
手順② 対象ブログデータ収集 — URLフィルター設定

- ブログの構成をみて、記事ファイルの命名規則を把握し、リスト等ではなく記事ファイルのみをダウンロードするよう設定します。
- 例えば、Amebloの場合には、次のような命名規則があります。
 - 新着記事一覧のファイル名・・・ /entrylist.html
 - 年月別の記事一覧のファイル名・・・/archiveentrylist-[yyyymm].html
 - 記事本体のファイル名・・・/entry-[数値].html
- 今回は記事本体のみを収集したいので、URLフィルタで[ドメイン]/entry-*.htmlを指定します。



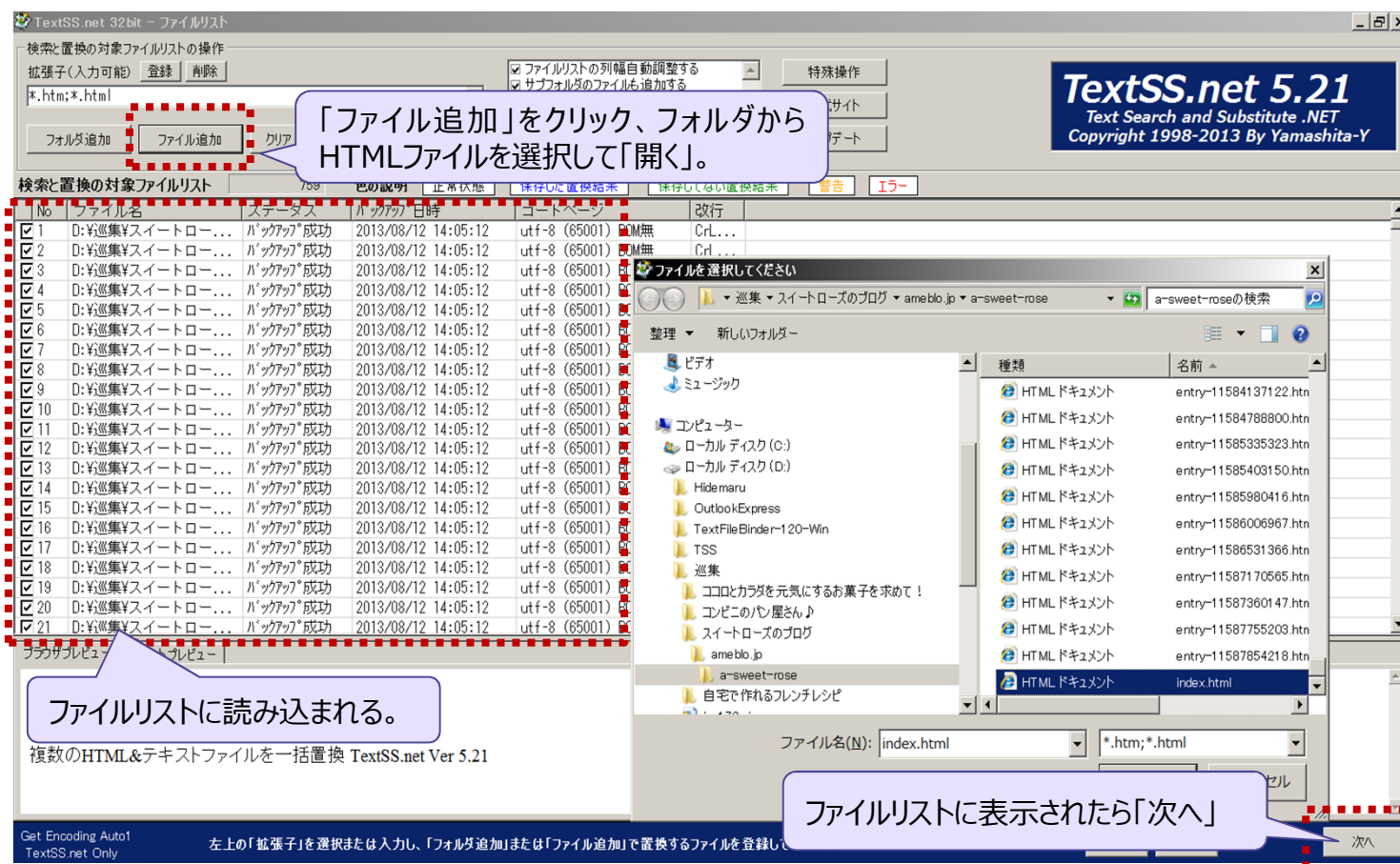
手順② 対象ブログデータ収集 — ダウンロード実行

- ダウンロード>ダウンロード実行をクリックすると、状況タブに進行状況が表示されます。
- ダウンロードが終了すると、オプションで指定したフォルダに、サイト名のフォルダができていることが確認できます。



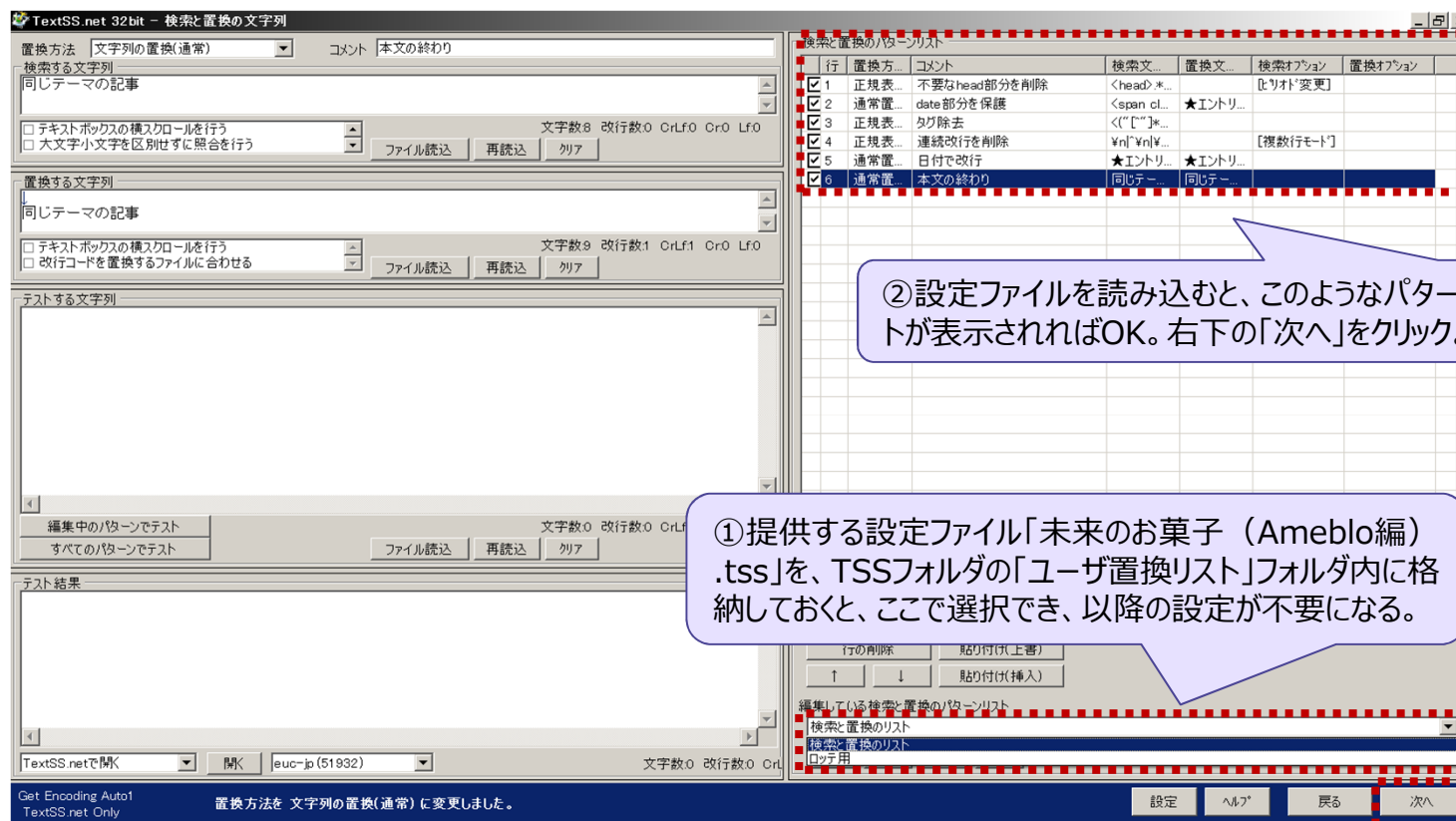
手順③ 収集した記事の加工 — TextSS.net.exeの起動

- 収集した記事はHTMLタグを含むファイルとなっています。TextSS.netを使い、不要なタグを除去したり、改行を加えたりすることで、未来表現やお菓子の検索をしやすい形に記事を加工します。
- TextSS.netの起動は、次のように行います。



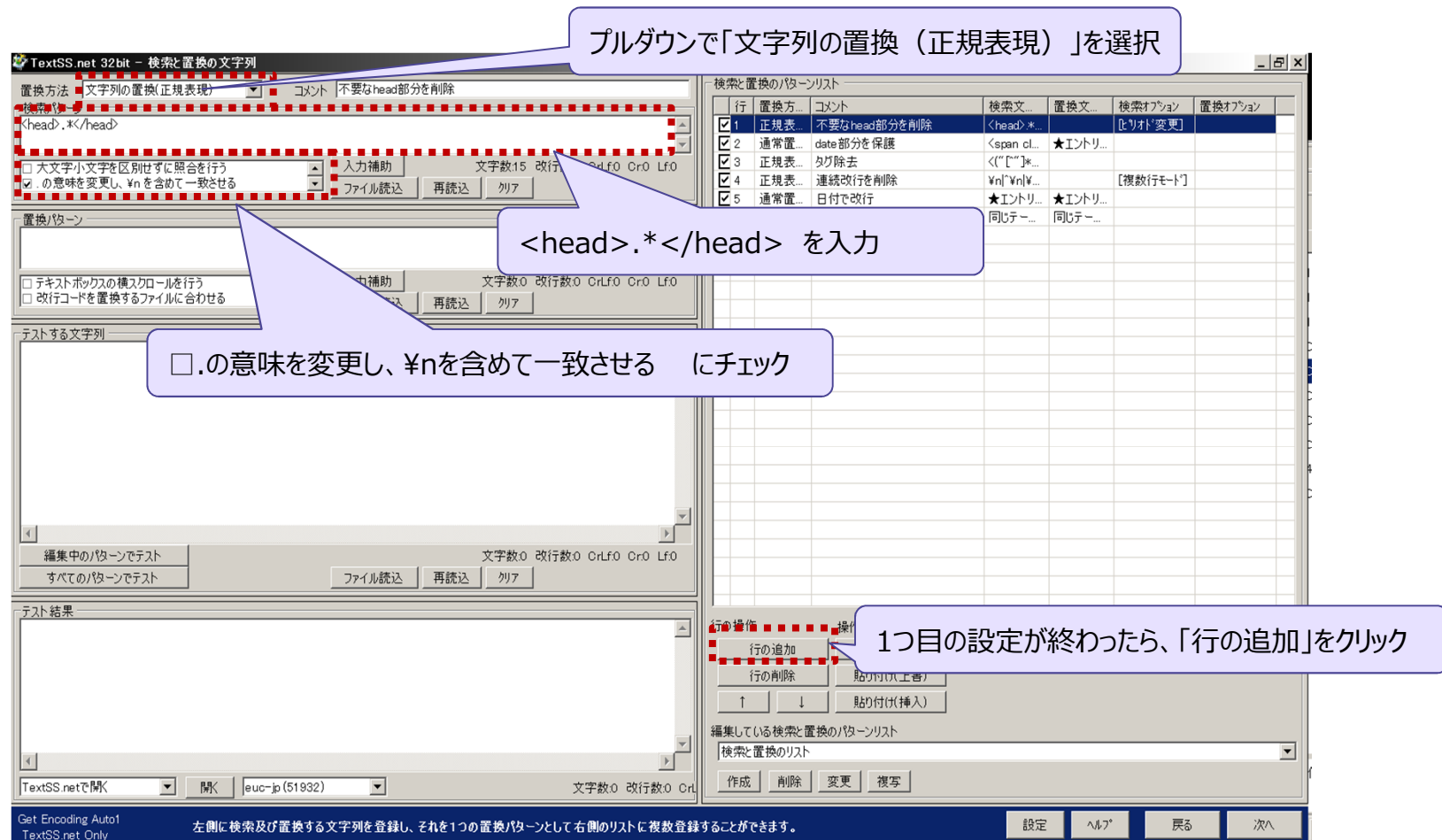
手順③ 収集した記事の加工 — TextSS.netへの設定ファイルの読み込み

- TextSS.netによって行う記事の加工処理は複数ありますが、1度使用した処理内容を「設定ファイル」として保存し、2回目以降に同じ処理を呼び出して実行することが可能です。
- 今回の取組みでは、P.21～26の設定内容について、「未来のお菓子（Ameblo編）.tss」として取りまとめてあります。記事の加工を行う際は、このファイルを次のように読み込むことで、必要な処理を素早く行うことができます。



手順③ 収集した記事の加工 — 設定1) 不要なヘッダ一部分の削除

- 収集した記事のうち、データとして必要なのは記事の本文部分のみにあたります。
- 本文前に必ず存在する「ヘッダー部分」を削除するために、次のような設定を行います。



手順③ 収集した記事の加工 — 設定2) 日付部分の保護

- 記事の加工時に削除するHTMLタグの中には、記事をエントリーした「日付」を示すタグも含まれます。この「日付」の情報は、後の工程で「1記事」の切れ目を判断したり、記事をソートしたりする際に使用するので、削除されないように保護する必要があります。
- 日付部分の保護の設定は、次のように行います。

プルダウンで「文字列の置換（通常）」を選択

★エントリー日付★

★エントリー日付★ と入力

行	置換方...	コメント	検索文...	置換文...	検索オプション	置換オプション
1	正規表...	不要なhead部分を削除	<head>*		[リセット]	
2	通常置...	date部分を保護	<span cl...	★エン...		
3	正規表...	タグ除去	<[""]*			
4	正規表...	連続改行を削除	\n \n \n			[複数行モード]
5	通常置...	日付で改行	★エン...	★エン...		

TextSS.netで開く

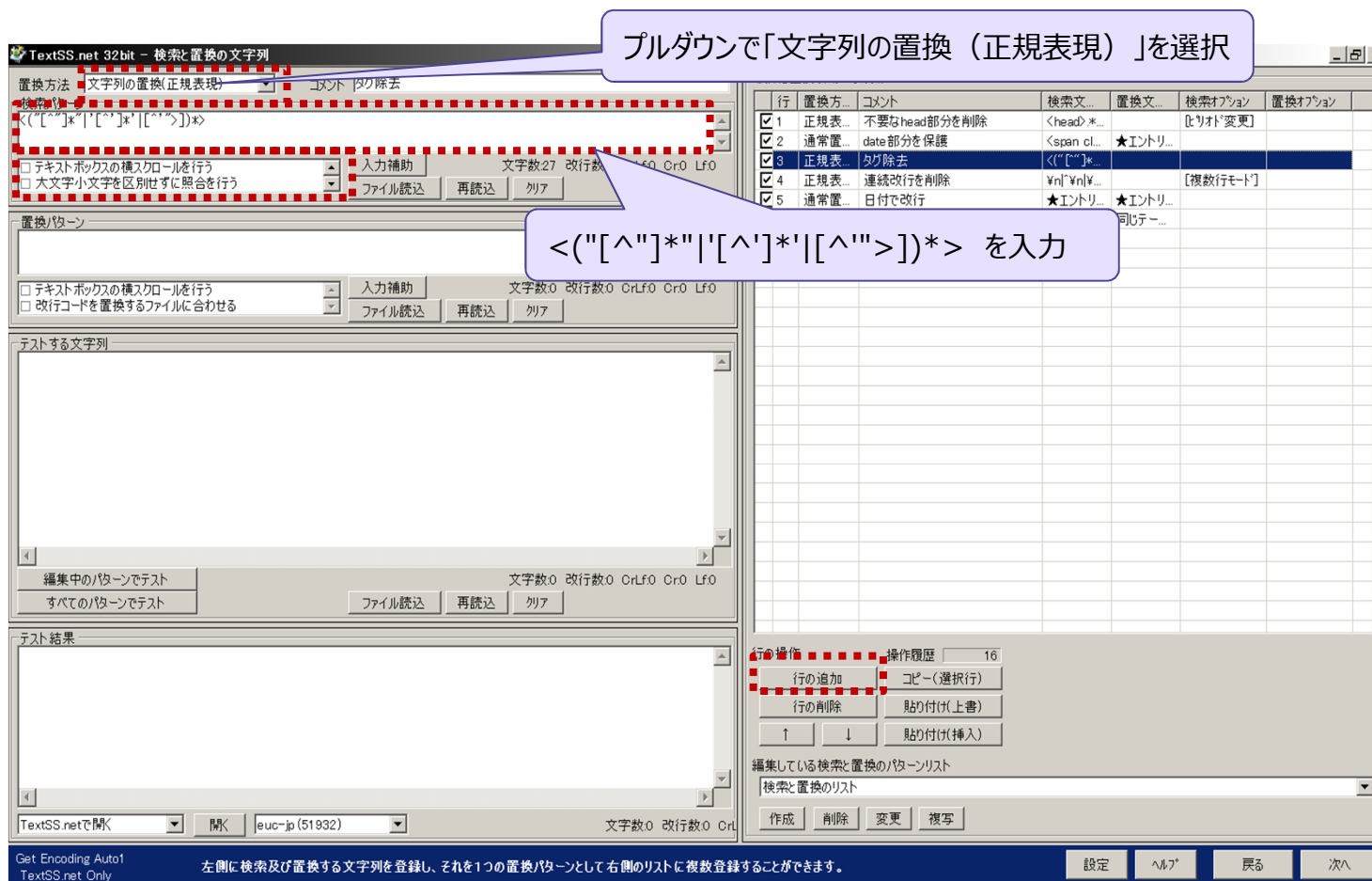
Get Encoding Auto1
TextSS.net Only

左側に検索及び置換する文字列を登録し、それを1つの置換パターンとして右側のリストに複数登録することができます。

設定 ヘルプ 戻る 次へ

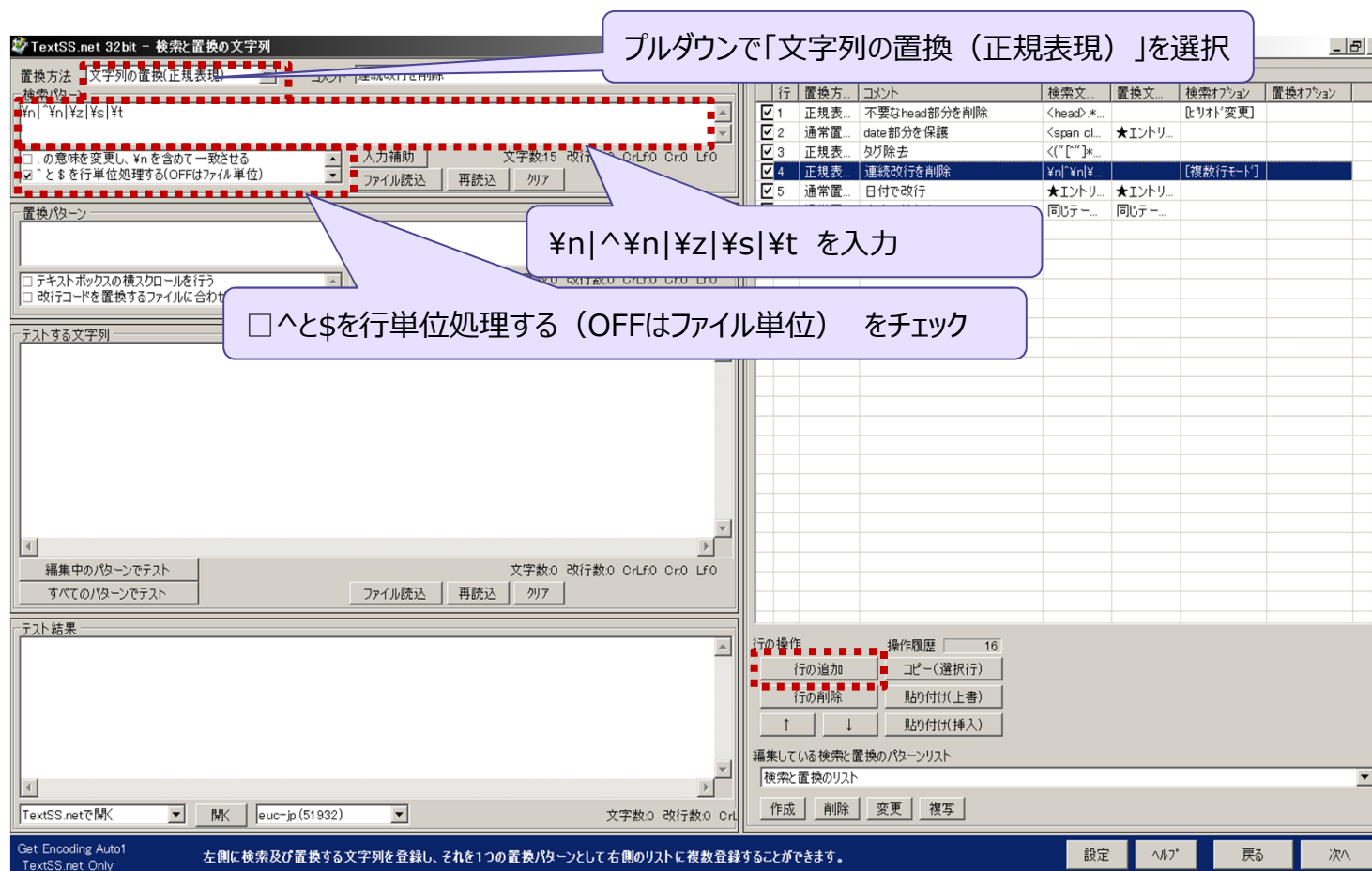
手順③ 収集した記事の加工 ー 設定3)HTMLタグの除去

- 収集した記事データから、不要なHTMLタグを削除する必要があります。
- HTMLタグの削除の設定は、次のように行います。



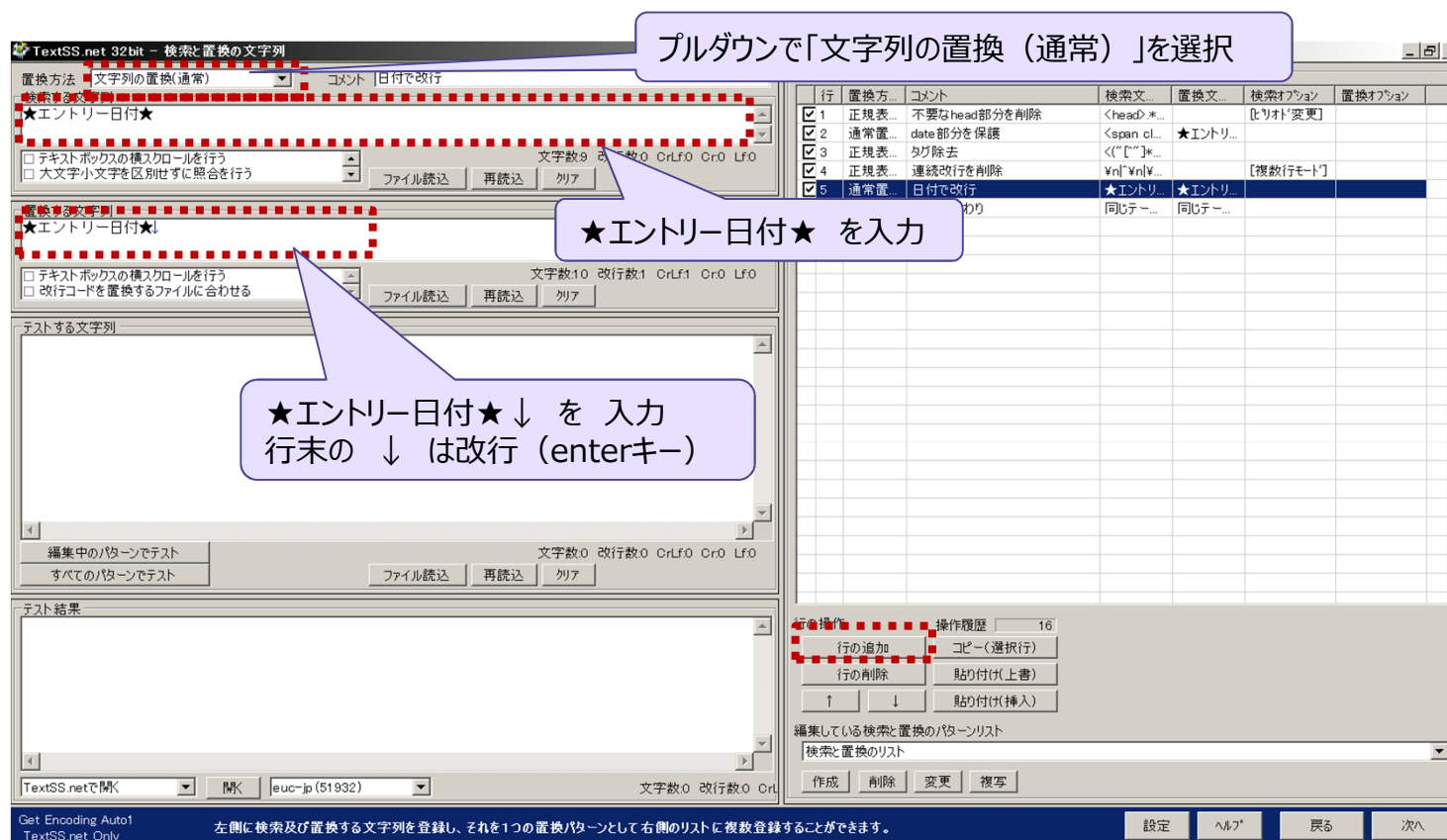
手順③ 収集した記事の加工 — 設定4)改行の削除

- 後の工程に向けて、収集した記事を「1記事1行」の状態に加工しておく必要があるため、本文中に含まれている改行を削除します。
- 改行の削除の設定は、次のように行います。



手順③ 収集した記事の加工 — 設定5) 日付前で改行

- 後の工程で、複数の記事データを1つのファイルにまとめますが、その際、1つ1つの記事の切れ目が分かるようになっている必要があります。記事のスタート地点は、あらかじめ保護してある「エントリー日付」の部分に存在するので、その部分に改行を挿入するように設定を行います。
- 日付前で改行を挿入する設定は、次のように行います。



手順③ 収集した記事の加工 — 設定6)本文の終わりで改行

- 後の工程で、複数の記事データを1つのファイルにまとめますが、その際、1つ1つの記事の切れ目が分かるようになっている必要があるので、本文の終わりに改行を挿入するように設定を行います。
- 本文の終わりに改行を挿入する設定は、次のように行います。

The screenshot shows the TextSS.net 32bit application window. The 'Replace Method' (置換方法) is set to 'Text String Replacement (Normal)' (文字列の置換(通常)). The 'Comment' (コメント) field contains 'End of Article' (本文の終わり). The 'Replace Text' (置換文) field contains '\n'. The 'Search Text' (検索文) field contains '本文の終わり'. The 'Replace Text' (置換文) field contains '\n'. The 'Search Options' (検索オプション) are set to 'Match Case' (大文字小文字を区別する) and 'Match Whole Word' (単語単位で一致を判定する). The 'Replace Options' (置換オプション) are set to 'Replace All' (すべて置換) and 'Replace in Place' (inplace 置換). The 'Test Text' (テストする文字列) field contains '本文の終わり'. The 'Test Result' (テスト結果) field is empty. The 'File' (ファイル) menu is open, showing 'Open' (開く), 'Save' (保存), 'Print' (印刷), 'Exit' (終了), and 'About' (ヘルプ). The 'Settings' (設定) button is highlighted in the bottom right corner.

プルダウンで「文字列の置換（通常）」を選択

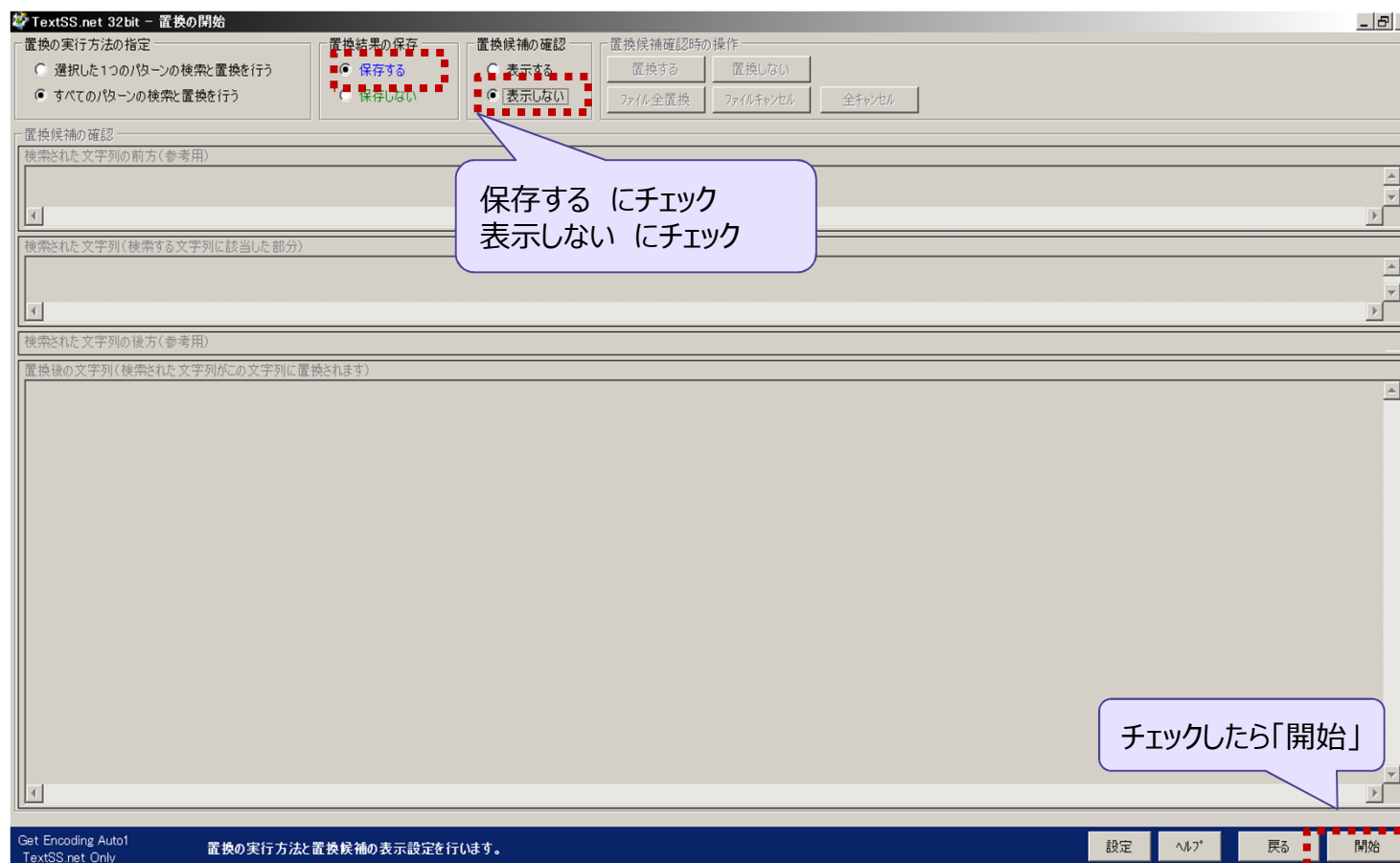
同じテーマの記事 を入力

↓
同じテーマの記事
を入力 行頭の↓は、改行(enterキー)

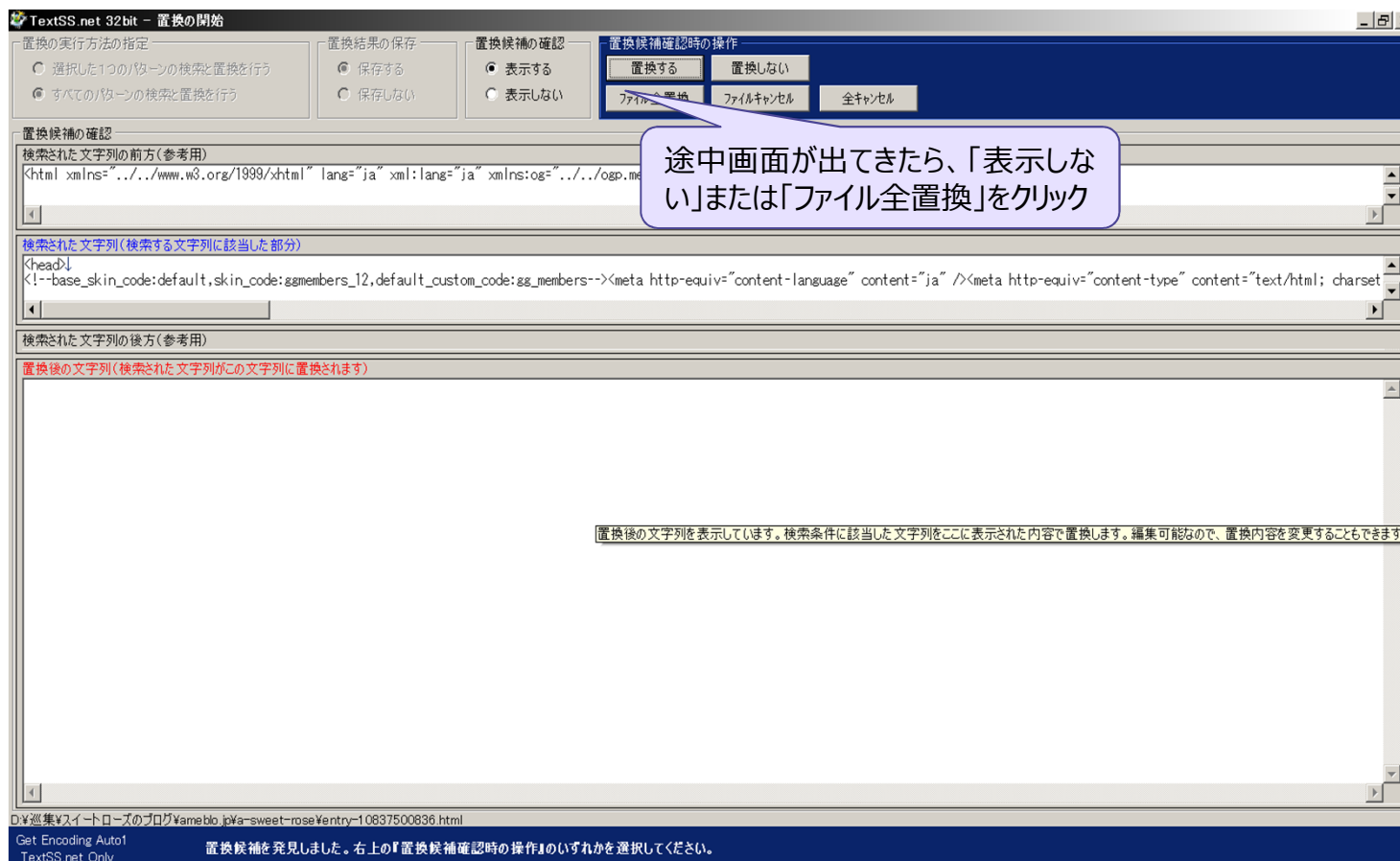
設定が終わったら「次へ」

手順③ 収集した記事の加工 — 置換の実行(1/4)

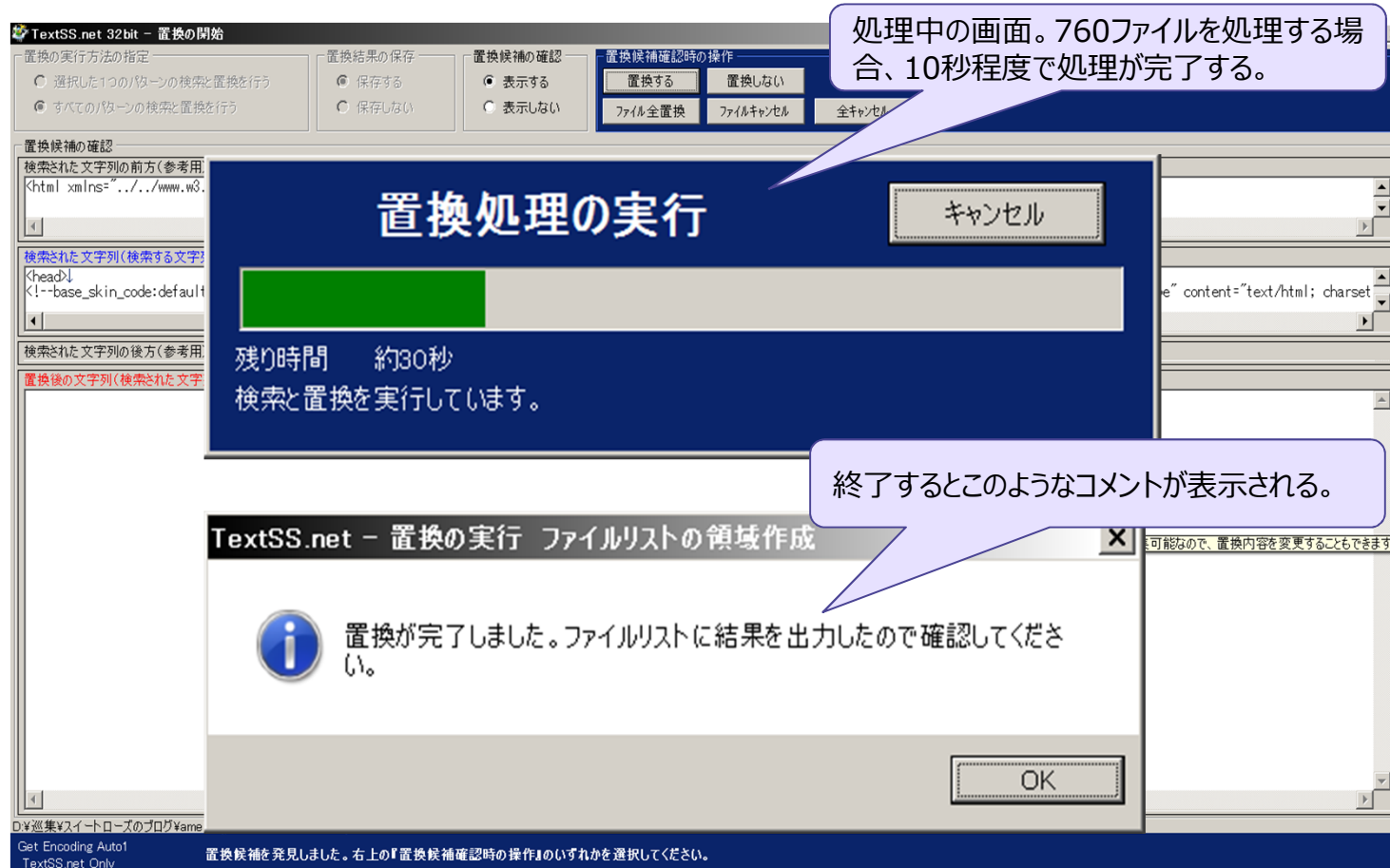
- 設定した処理内容に従い、記事の加工を行います。
- 加工の手順は次の通りです。



手順③ 収集した記事の加工 — 置換の実行(2/4)



手順③ 収集した記事の加工 — 置換の実行(3/4)



手順③ 収集した記事の加工 — 置換の実行(4/4)

The screenshot displays the TextSS.net 5.21 application window. The main window is titled "TextSS.net 5.21" and shows a list of files being processed. A callout box points to the "復元" (Restore) button, stating: "元のHTMLファイルは、不要な箇所が削除されたため、サイズが小さくなっている。置換の実行と同時に、元のファイルのバックアップが取られるので、失敗しても「復元」ボタンで元に戻る。" (The original HTML file is smaller because unnecessary parts were removed. Since a backup of the original file is taken simultaneously with the replacement, even if it fails, you can return to the original by clicking the "Restore" button.)

Another callout box points to the first row of the file list, stating: "最初のファイルリストに処理結果が表示される。" (The processing result is displayed in the first file list.)

The file list shows the following data:

No	ファイル名	ステータス	バックアップ日時	コードページ	改行
1	entry-10933153487.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
2	entry-10933153487.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
3	entry-10934502026.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
4	entry-10934502026.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
5	entry-10935404876.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
6	entry-10935404876.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
7	entry-10936402190.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
8	entry-10936402190.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
9	entry-10937762798.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
10	entry-10937762798.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
11	entry-10938218262.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
12	entry-10938218262.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
13	entry-10938234521.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
14	entry-10938234521.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
15	entry-10939631389.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
16	entry-10939631389.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
17	entry-10939942130.html	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf
18	entry-10939942130.html.bak	置換完全成功	2013/08/12 14:05:12	utf-8 (85001) BOM無	CrLf

The bottom status bar indicates: "Get Encoding Auto1 TextSS.net Only 置換が完了しました。ファイルリストに結果を出力したので確認してください。" (Replacement is complete. Please check the results in the file list.)

手順③ 収集した記事の加工 — テキストの結果

- 加工処理が完了したファイルをテキストエディタ（秀丸など）で開くと、次のような形式になっています。

1 記事分のHTMLファイルは、このように 3 行構成（改行は 2 しかない）になっており、改行で囲まれた真ん中の日付で始まる行が本文。

The screenshot shows a web browser window with a Japanese blog post. The browser's address bar displays the URL: [D:\Y集集サイトエントリのブログYamabito.jp-sweet-rooseyentry-10837500836.html](http://d.hatena.ne.jp/yamabito-jp/sweet-rooseyentry-10837500836.html). The browser's menu bar includes options like 'ファイル(E)', '編集(E)', '表示(V)', '検索(S)', 'ウインドウ(W)', 'マクロ(M)', and 'その他(O)'. The browser's status bar at the bottom shows '秀九デバッグ', '下候補', '単語をこ', '分割ウインドウ切り抜き', '貼付け', 'タグジャンプ', 'アライン解析', '行番号表示', 'Unicode(UTF-8)', and '挿入モード'.

The blog post's header features a navigation bar with links: 'D.Y集集サイトエントリのブログYamabito.jp-sweet-rooseyentry-10837500836.html' and '秀九'. Below the navigation bar is a search bar and a list of links: 'ファイル(E)', '編集(E)', '表示(V)', '検索(S)', 'ウインドウ(W)', 'マクロ(M)', and 'その他(O)'. The main content area displays a list of blog posts, each with a date and a title. The first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The second post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The third post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fourth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The seventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The eighth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The ninth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The tenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The eleventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twelfth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirteenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fourteenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifteenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixteenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The seventeenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The eighteenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The nineteenth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twentieth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-second post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-third post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-fourth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-fifth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-sixth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-seventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-eighth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The twenty-ninth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirtieth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-second post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-third post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-fourth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-fifth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-sixth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-seventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-eighth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The thirty-ninth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fortieth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-second post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-third post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-fourth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-fifth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-sixth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-seventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-eighth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The forty-ninth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fiftieth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-second post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-third post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-fourth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-fifth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-sixth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-seventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-eighth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The fifty-ninth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixtieth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-second post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-third post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-fourth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-fifth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-sixth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-seventh post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-eighth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The sixty-ninth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The seventieth post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:22:22'. The seventy-first post is dated '2011-03-21 22:22:22' and has the title '2011-03-21 22:

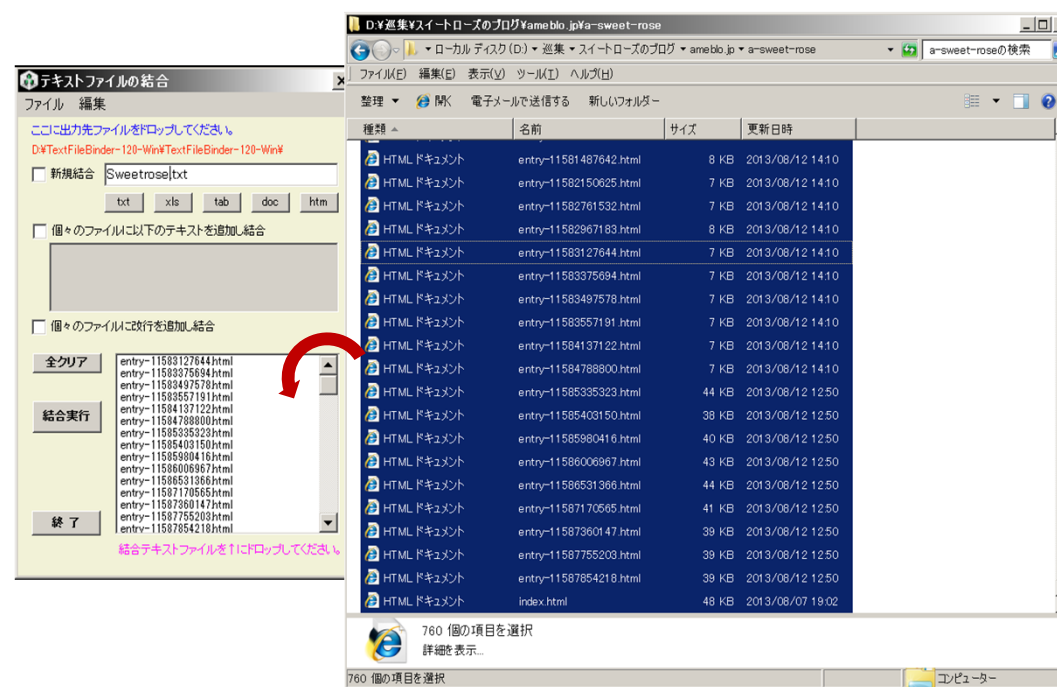
手順③ 収集した記事の加工 — HTMLドキュメントの結合

- 次の工程では、TextSS.netでは処理しきれなかった不要な文章データを削除する作業を行います。ここでは、TextSS.netを使った処理のように、1度に複数のファイルに対して処理を施すことができないので、1ファイルに1回の処理を施す必要があります。
- しかし、処理の対象となるテキストファイルは膨大な数になるため、1つ1つのテキストファイルを全て処理するのは非現実的です。そのため、テキストデータ結合ツール「Text Binder」(<http://www.asahi-net.or.jp/~gj3t-tki/REALbasicAPL/TextFileBinder.html>)を使用し、複数の記事ファイルを統合し、総ファイル数を減らします。
- 一度に大量の記事を選択しすぎると、結合に失敗したり、この後の手順に失敗することがあるので、選択する記事数には注意が必要です。以降の操作をスムーズに進めるには、100記事程度ずつを目安に、1つのファイルに統合してください。

- 「Text Binder」を起動し、ファイルの出力先とファイル名を指定します。

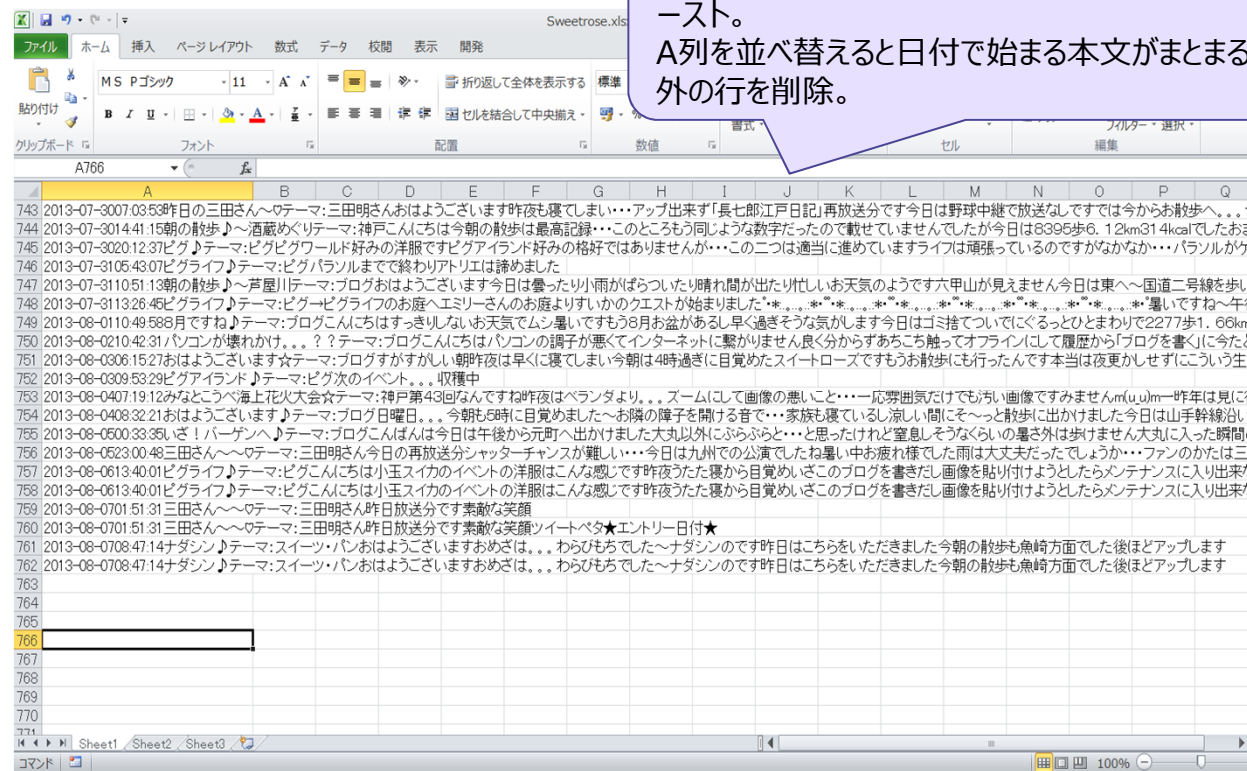
- フリーツール「巡集」で収集した記事ファイルを複数選択し、「Text Binder」の指定のWindowにドロップします。

- 「結合実行」ボタンをクリックすると、選択した記事データが1つのテキストデータに結合されます。



手順③ 収集した記事の加工 — エクセルでの並べ替えと不要行の削除

- 統合されたテキストファイルから不要な文章を削除する作業はExcelで行います。
- 統合されたテキストファイルを、テキストエディタで開いてコピーし、Excelに貼り付け、ソートを行います。これまでの処理の結果、最終的に必要となるデータの冒頭は全て「日付」で始まるように加工を行っているため、必要なデータがまとまった状態を作ることができます。
- 日付で始まるセル以外は不要なデータなので、削除してください。



手順④ お菓子・未来表現検索 — 正規表現検索ツールのインストール(1/2)

- 未来表現の検索を容易に行うために、Excel用の正規表現検索ツールを利用します。
- <http://srcedit.pekori.jp/tool/excelre.html>から、「Excel用正規表現検索ダイアログアドイン」をダウンロードします。
- インストールの手順(<http://srcedit.pekori.jp/tool/reinst.html>)に従い、ツールを使用可能な状態にします。

◆ダウンロード

正規表現検索
Excel用正規表現検索ダイアログアドイン

紹介

～Excel用正規表現検索・置換ダイアログ～

正規表現による検索・置換ダイアログを表示するExcelアドインです。インストールするとExcelで正規表現を使用した柔軟な検索が行えます。インターフェースがExcel標準の検索に近いので使い易いと思います。また正規表現に関するワークシート関数を使用できるようになります。

オートシェイプの検索、置換、Grep、Grep置換も可能です。

オートフィルタに正規表現を使用する代替として列のデータに対して正規表現にマッチするデータを抽出する機能もあります。

正規表現でExcelファイルのGrep、Grep置換も可能です。

起動ショートカットキーのカスタマイズができます。

アドインのインストール、アンインストールを自動で行うので導入に手間がかかりません。

Excel2000以上で動作します。

フリーウェアです。

このサイトで紹介されました。

ソフトバンク クリエイト株式会社 PC Japan 2009年10月号で紹介されました。

晋遊舎 Mr.PC 2009年Vol.5で紹介されました。

晋遊舎 Windows100% 2010年07月号で紹介されました。

晋遊舎 Mr.PC 2011年09月号で紹介されました。

晋遊舎 Windows100% 2011年10月号で紹介されました。

晋遊舎 Mr.PC 2011年12月号で紹介されました。

晋遊舎 Windows100% 2013年02月号で紹介されました。

晋遊舎 IP! 2013年03月号で紹介されました。

スクリーンショット

検索ダイアログ
置換ダイアログ

ダウンロード

正規表現検索2.46 (2013/07/27更新)

ここをクリックすると試用できます。

マクロが使用可能なExcel2000以上が必要で、クリックした後、開くを押して、マクロを有効にします。

するとアドインのインストールをせずに一時的に(フォームを開くまで)試用できます。

試用中は、一部の機能が使用できなくなっています。

最新版の変更点

- 細かい修正。
- 細かいバグ修正。

◆インストール手順①

圧縮ファイルを解凍後「excelre.xla」をダブルクリックします。

名前	更新日時	種類	サイズ
auto.xla	2011/11/27 17:48	Microsoft Office...	35 KB
excelre.xla	2012/08/12 15:13	Microsoft Office...	1,285 KB
FunCustomize.dll	2003/03/19 22:54	アプリケーション...	59 KB
readme.txt	2012/08/12 14:56	テキスト文書	34 KB
readme_e.txt	2012/08/12 14:56	テキスト文書	36 KB

マクロを有効にして開きます。

セキュリティ警告

“C:\... \WexcelreWexcelre.xla” はマクロを含んでいます。

マクロにはウイルスが含まれている場合があります。通常、マクロを無効にすると安全ですが、マクロが適正な場合、機能が使えなくなります。

マクロを無効にする(D) マクロを有効にする(E) 詳細(M)

言語選択で日本語を選択するので「いいえ」を選びます。

Microsoft Excel

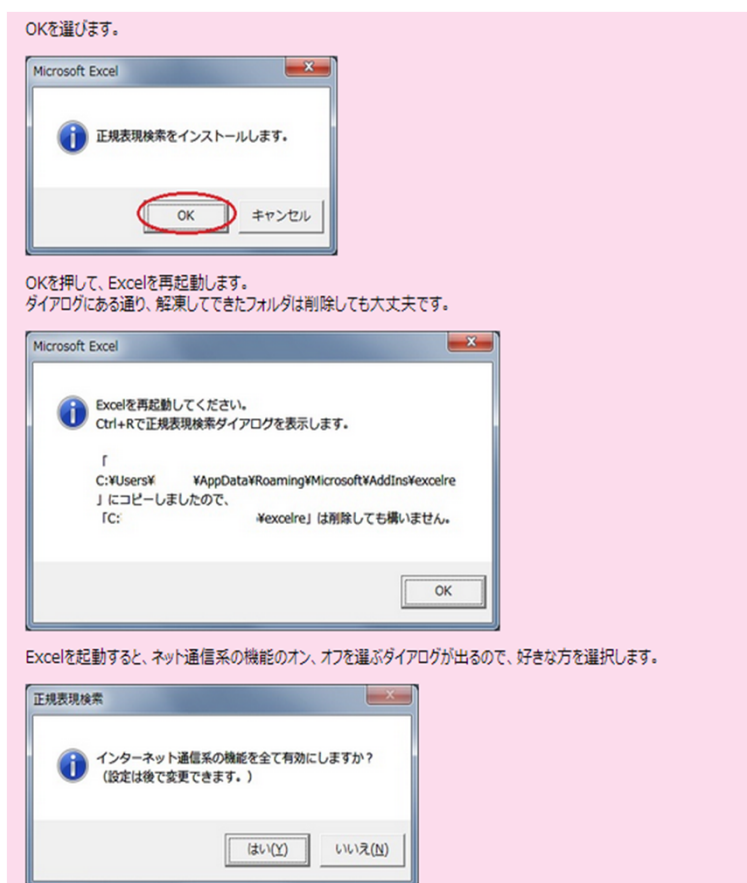
Language? Yes then English, No then Japanese.

はい(Y) いいえ(N)

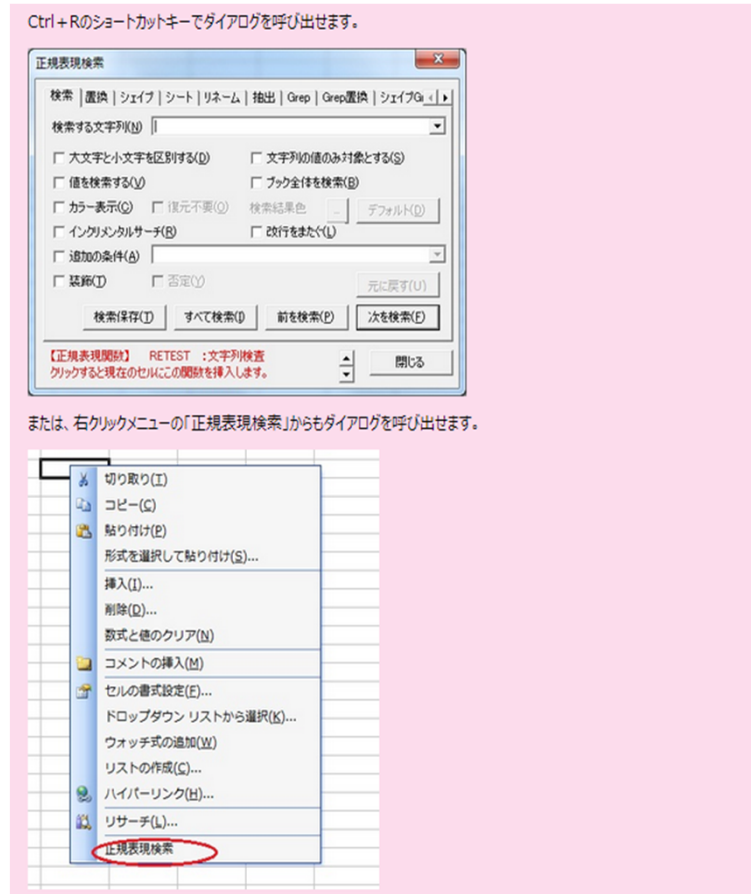
手順④ お菓子・未来表現検索 — 正規表現検索ツールのインストール(2/2)

- <http://srcedit.pekori.jp/tool/excelre.html>から、「Excel用正規表現検索ダイアログアドイン」をダウンロードします。
- インストールの手順(<http://srcedit.pekori.jp/tool/reinst.html>)に従い、ツールを使用可能な状態にします

◆インストール手順②

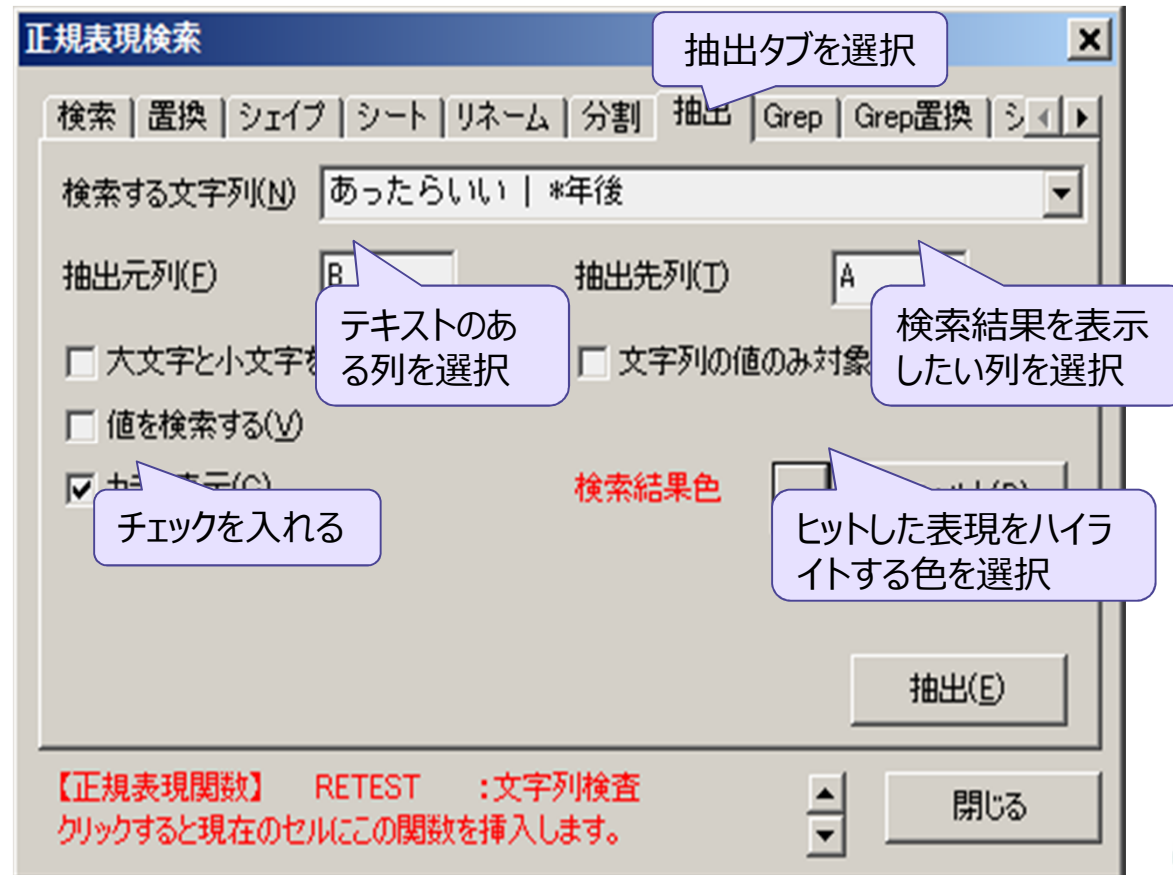


◆インストール手順③



手順④ お菓子・未来表現検索 — 未来表現のあるテキストの抽出

- 「Excel用正規表現検索ダイアログアドイン」のインストールを完了したら、Ctrl+Rで下図のウィンドウを呼び出すことができます。
- 下図の通りにすべての項目をセットしてから「抽出(E)」をクリックすると、正規表現を用いてすべてのテキストデータの検索が行われ、結果がエクセルシートに反映されます。(結果のイメージは次のページに示します。)



手順④ お菓子・未来表現検索 — 検索の結果

■ 検索後の結果表示のイメージです。

	A	B	C	D	E	F	G
1	1						
2	1	2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
3	1	2013年02月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
4	1	2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
5	1	2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
6	1	2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
7	1	2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
8	1	2013年02月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
9	1	2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
10		2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
11		2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
12		2013年03月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
13	1	2013年04月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
14		2013年04月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	
15	1	2013年04月のブログ	*スイート	ローズ*	幸せのひととき♪	本文	

画像はTextSS.net使用以前の
手順の結果なので、
処理済みのデータが得られた段階で
張り替えます。

テキストデータのあるセルを展開すると、下記のようにヒット
したキーワードがハイライト表示されています。

識は、非常に大事な学生時の勉強でもあるもので、「教育」関連、「大学」の内容とか、より充実したカリキュラムにすることで、**未来**の繁栄につながる人材が育って、社会に羽ばたくと、これまた素晴らしいアイデアでヒット商品が開発されたりなんかして、繁栄は繁栄を呼ぶ。「ちょい高」傾向というか、「ちょいリッチ」傾向って、「ちょっと幸せ。」の時間を持つのにも**買っているかも**ね。（写真のコーヒーは、アロマリッチに6種類のコーヒーがセレクトされたもの）スーパーで見かけた「サンラヴィアン」のブロックケーキ（モンドセレクション金賞受賞商品）でも、いろいろな種類のものを詰め合わせたものがあつたけど、